

GRASP-DNA: A Web Application to Screen Prokaryotic Genomes for Specific DNA-Binding Sites and Repeat Motifs

Christophe H. Schilling^{1*}, Lance Held¹, Mike Torre² and Milton H. Saier, Jr.²

¹Department of Bioengineering University of California, San Diego 9500 Gilman Dr., 0412 La Jolla, CA 92093-0412, USA

²Department of Biology University of California, San Diego 9500 Gilman Dr., 0116 La Jolla, CA 92093-0116, USA

Abstract

The ability to control multiple genes at the transcriptional level often relies on the existence of short stretches of well-defined DNA sequences, to which regulatory proteins and transcription factors bind. In this article we present a freely accessible web-based application (GRASP-DNA), that can be used to screen prokaryotic genomes for putative DNA-binding sites of a particular transcription factor or DNA-binding molecule. This application utilizes existing theories, such as information and statistical-mechanical theories, for the calculation of positive weight matrices generated from block aligned binding sites. Using these position weight matrices entire prokaryotic genomes are screened to identify sites that display a high level of sequence similarity to existing binding sites. This application can be used in combination with high-throughput technologies for gene expression analysis and binding site characterization to assist in the elucidation of global regulatory networks.

Introduction

From genome sequencing and annotation efforts lists of thousands of genes are constructed, potentially providing a complete description of the structure and content of an organism's genome. With this information in hand the challenge shifts to understanding how these genes are collectively utilized to orchestrate multi-genetic cellular functions. What are the rules and regulatory logic implemented by the cell to selectively express subsets of genes in a coordinated fashion?

In many cases the transcription of a gene is controlled by short stretches (5-25 base pairs) of well-defined DNA sequences to which regulatory proteins and transcription factors bind. Sequence specific protein-DNA binding can affect transcription of a gene in both a negative and a positive fashion based on parameters including the location

of factor binding relative to the promoter binding site and the translational start site of a gene. Many regulatory proteins are pleiotropic, individually providing for the global regulation of a large number of genes, which are consequently referred to as members of a single regulon. Such regulators allow for coordinated cellular responses to environmental stresses and challenges. Among such pleiotrophic regulators in *Escherichia coli* are the cyclic AMP receptor (CRP) and catabolite repressor/activator (Cra) proteins that control carbon metabolism (Saier, 1989; Saier and Ramseier, 1996), the NtrC regulator that controls nitrogen metabolism (Magasanik, 1993), and the PhoB regulator that influences phosphorus metabolism (Wanner, 1993).

With the availability of an increasing number of complete genome sequences, it is now possible to perform exhaustive computational screening of genomes for putative regulatory binding sites based on the sequence of presumed or biochemically-established binding sites. Such computational analysis can be used as a starting point for the elucidation of novel binding sites, and can be complemented by experimental technologies including high-throughput expression profiling. Numerous computational programs are available to allow researchers to use pre-constructed matrices generated from a list of known binding sites to search a particular sequence string and predict proteins that are likely to bind to the given sequence (Frech *et al.*, 1997a). Programs are also available for the generation of consensus sequences and position weight matrices that can be used to search sequences for additional binding sites (Frech *et al.*, 1997b). Other programs perform sequence searches using pre-constructed matrices from databases of transcription factor binding sites (Heinemeyer *et al.*, 1999; Salgado *et al.*, 1999).

We were interested in developing a web-based application for the experimentally inclined user to assist in the identification of potential binding sites for a particular DNA binding protein of choice in a complete genome. There are many examples where known DNA binding sites for a protein have been used to computationally screen for additional sites and further elucidate the interaction of various genetic regulatory networks (Lewis, 1994; Tronche *et al.*, 1997; Thieffry *et al.*, 1998; Wasserman and Fickett, 1998). In a recent study, 55 of the 240 candidate *E. coli* DNA-binding proteins were used to generate binding site matrices for searching the entire *E. coli* genome to predict novel binding sites (Robison *et al.*, 1998); a database of these matrices and search results was constructed.

Here we introduce a freely accessible web-based application (GRASP-DNA – accessible at <http://www-bioeng.ucsd.edu/~grasp>) developed to allow researchers to utilize a list of known regulatory binding sites for a specific protein to screen prokaryotic genomes. Clients simply

Received May 22, 2000; accepted May 25, 2000. *For correspondence. Email cschilli@bioeng.ucsd.edu; Tel. (858) 822-1144; Fax. (858) 822-3120.

provide a list of block-aligned binding sites for the regulatory protein of their choice, and GRASP-DNA will calculate and display a position weight matrix that is then used to screen and rank potential binding sites in a genome of choice.

Detecting Potential Regulatory Sites

Computational approaches used to search for novel regulatory sites involve the use of either crude consensus sequences or more precise position weight matrices generated from a series of aligned binding sites (Stormo, 1988; Stormo, 1990). In a position weight matrix a specific value is assigned to each nucleotide for every position in the sequence indicating the "contribution" of the particular base to the specificity of the binding site. The overall score of a DNA sequence window of equal length to the binding sites is the sum of all the values for each position. The score then reflects the binding affinity or specificity of the sequence for the DNA-binding protein and allows different

sequences to be quantitatively compared as potential binding sites.

A position weight matrix can be generated from an occurrence or frequency table using a choice of different mathematical approaches. Each approach attempts to provide a score, s_{bp} , for each nucleotide base (b) in position (p) of the aligned binding sites. For any target sequence of the same length as the binding sites, an overall score (S) can be calculated as a strict linear combination of the scores assigned to each base in the target sequence as shown below:

$$S = \sum_{p=1}^L s_{bp} \quad \text{Equation 1}$$

where L is the length of the binding site. In GRASP-DNA we offer the user a selection of two alternative approaches to calculate s_{bp} and a complete position weight matrix based

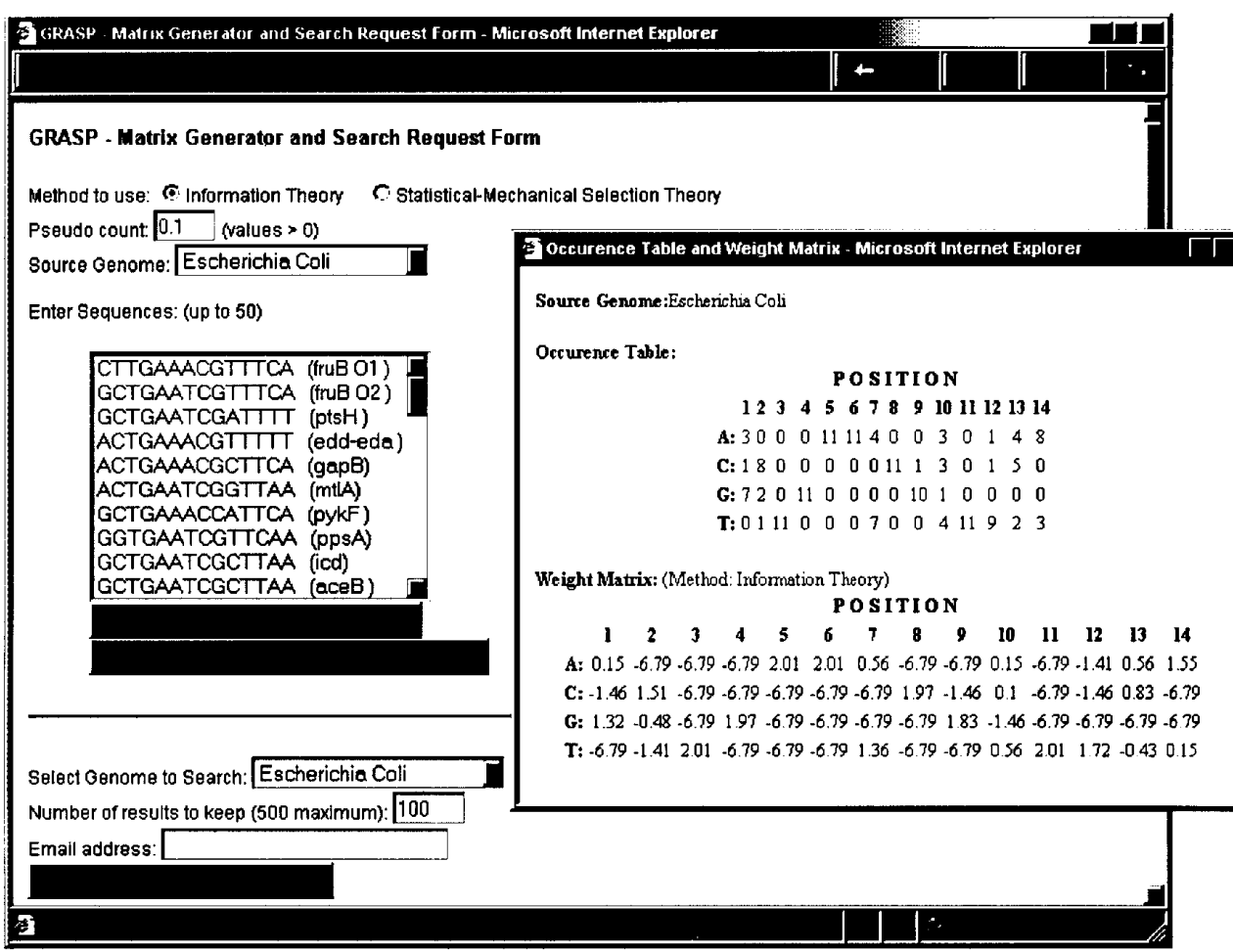


Figure 1. Screen capture of the GRASP-DNA search request form and the accompanying position weight matrix generated from the input parameters. In this particular case eleven 14-bp DNA sequences are entered which correspond to FruR binding sites from *E. coli*, and information theory is selected as the matrix calculation method.

on either information theory or statistical-mechanical theory. Both of these approaches have been successfully implemented in screening for putative regulatory binding sites, and we do not advocate one approach versus another as both will generate comparable results to be further processed and complemented by experimental findings.

In the next few paragraphs the details of these two approaches are discussed for the interested reader.

Information theory attempts to determine how constrained the choice of a base is at each position within a binding site, and then estimates the amount of information that is encoded in a local sequence window as the linear

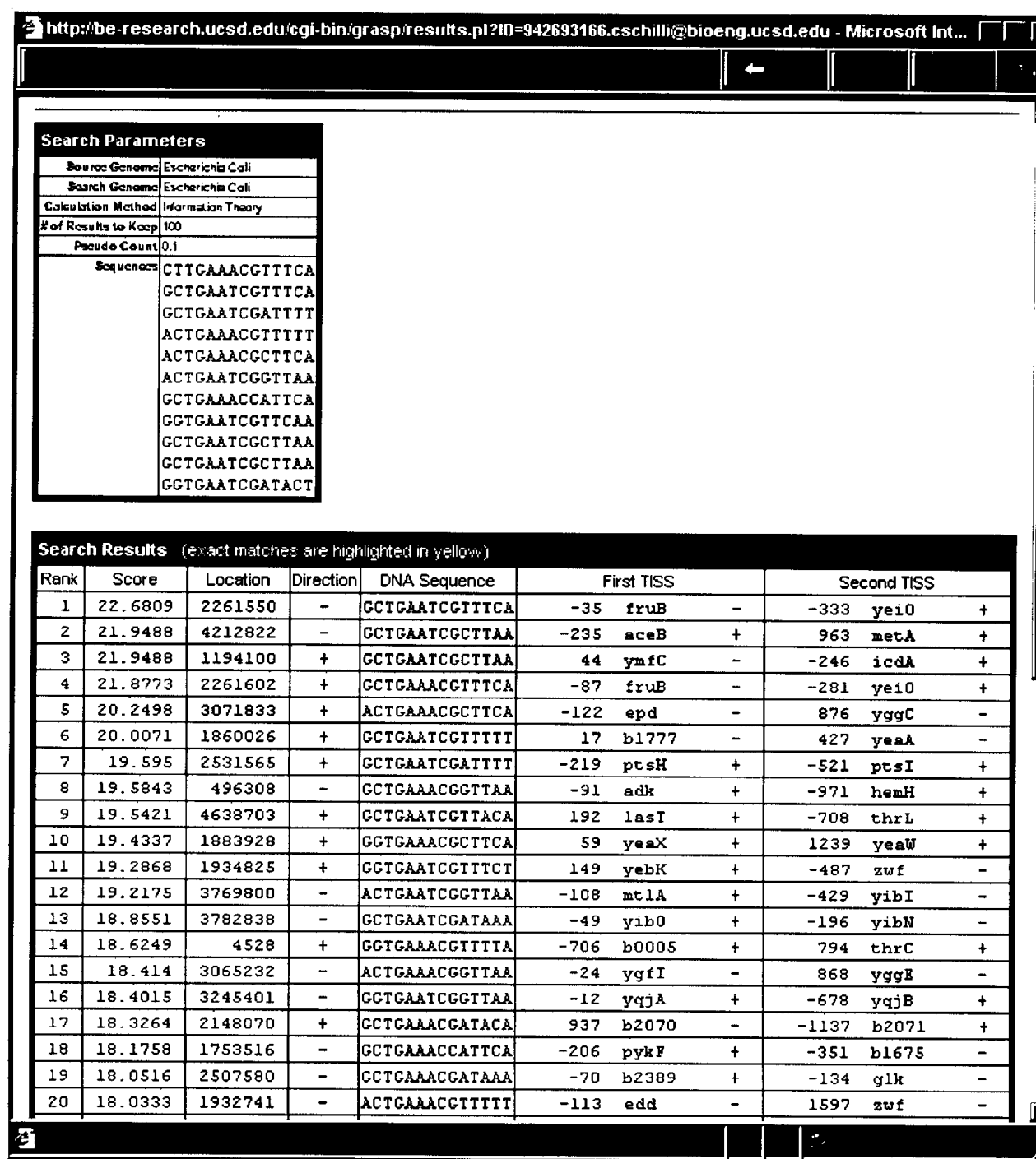


Figure 2. Sample results page for a screen of the *E. coli* genome for additional FruR binding sites using the position weight matrix calculated in Figure 1. Sites that match the exact sequence of one of the input sequences are highlighted in yellow in the online application. The table of search results can be sorted online by any of the corresponding fields in the table.

summation of the scores at each position (Stormo, 1988; Schneider, 1997). The standard log transformation equation used to calculate the individual position scores is shown below:

$$s_{bp} = \log_2 \left[\frac{f_{bp}}{q_b} \right] \quad \text{Equation 2}$$

where f_{bp} is equal to the observed frequency of nucleotide b at position p in the series of aligned sequences, and q_b is the frequency of finding nucleotide b within the source genome of the binding sites (determined from genomic nucleotide composition). Since only a limited number of sites are typically used to generate the matrix, the frequency at which some nucleotides are observed in a particular position may equal zero. This creates an undefined score in equation 2. A simple solution to this problem is to introduce what is called a 'pseudo-count' parameter (c) to ensure that the fraction in equation 2 will never equal zero in the case of an unobserved nucleotide (Claverie and Audic, 1996). Rather than traditionally equaling the number of occurrences (O_{bp}) divided by the number of sites (N), the frequency variable f_{bp} can be calculated using the following equation:

$$f_{bp} = \frac{O_{bp} + c \cdot q_b}{N + c} \quad \text{Equation 3}$$

For detailed information on the statistical relevance of search results based on this form of position weight matrix generation see (Claverie, 1994).

It can be seen that the total information content of a sequence will generally be positive if there is more information in the sequence than would be expected at random. Furthermore, the higher the total score, (S), becomes then the greater the similarity between the particular sequence and the list of sites used to perform the search.

An entirely alternative approach to calculating the position scores is provided by statistical-mechanical theory, which generates scores that have been formally correlated

to experimentally characterized binding constants (Berg and von Hippel, 1987). Rather than calculating the position score as a function of the occurrence frequency of each base and the random base probability (q_b), the score is based on the ratio of occurrence frequency of each base to the occurrence frequency of the most commonly occurring base at the same position (O_{Hp}). The position scores are then calculated as shown below:

$$s_{bp} = \ln \left[\frac{O_{bp} + c}{O_{Hp} + c} \right] \quad \text{Equation 4}$$

Once again a pseudo-count value is introduced to avoid an undefined log transformation. The maximum score for a target sequence will be zero and will only occur when each of the nucleotides is equal to the nucleotide that occurs most frequently at that specific position. The less similar a site is to the initial set of known sites used to generate the position weight matrix, the more negative the total score will become. For both information theory and statistical-mechanical theory, a decrease in the pseudo-count will increase the contribution of the more frequent nucleotides with respect to those that appear less frequently, effectively increasing the overall stringency.

Performing a Genome Screen for Putative Binding Sites

Once logged onto the GRASP-DNA system the researcher can enter in a list of nucleotide sequences for a given regulatory binding site. A position weight matrix will then be calculated based on the specification of the calculation method, pseudo-count, and source genome (only used for information theory). For both calculation methods, suggested default values of the pseudo-count are provided to assist users who may be unfamiliar with the underlying mathematics. Thus, in its simplest form, the GRASP-DNA system can be used as an educational tool to gain a better understanding of the composition and calculation of weight matrices. These matrices are important mathematical constructs in the representation of promoter elements,

Table 1. The first six genes listed below (derived from Ramseier *et al.*, 1993) were used to construct a position weight matrix to search the *E. coli* genome for putative FruR binding sites. Both information theory and statistical-mechanical theory were used to perform the search with the pseudo-count set to equal a standard default value of 0.1 in both cases. The rankings of these 6 sites among all the possible positions scored in the genome (9,278,442 fourteen base pair sequence windows) are shown along with the ranking of 5 other binding sites for FruR that were subsequently shown to control transcription of the genes listed (see Ramseier *et al.*, 1995).

Gene (Species)	Operator Sequence	Location Ranking	
		Information Theory	Statistical-Mech. Theory
<i>fruB</i> O1 (Eco/Sty)	CTTGAAACGTTTCA	14	19
<i>fruB</i> O2 (Eco/Sty)	GCTGAATCGTTTCA	3	2
<i>ptsH</i> (Eco)	GCTGAATCGATTTT	9	9
<i>ppsA</i> (Eco)	GGTGAATCGTTCAA	6	6
<i>icd</i> (Eco)	GCTGAATCGCTTAA	2	1
<i>aceB</i> (Eco)	GCTGAATCGCTTAA	1	3
<i>edd-eda</i> (Eco)	ACTGAAACGTTTTT	228	255
<i>gapB</i> (Eco)	ACTGAAACGCTTCA	65	62
<i>mtlA</i> (Eco)	ACTGAATCGTTTAA	294	171
<i>pykF</i> (Eco)	GCTGAAACCATTTCA	137	140
<i>pckA</i> (Eco)	GGTGAATCGATACT	257	268

splice sites, and other DNA sequence and protein motifs.

Once a satisfactory position weight matrix is constructed, the researcher can screen for putative binding sites in any one of seven prokaryotic genomes currently available for screening in the GRASP-DNA system. The genomes include *Bacillus subtilis*, *Escherichia coli*, *Haemophilus influenzae*, *Mycoplasma genitalium*, *Helicobacter pylori*, *Mycobacterium tuberculosis*, and *Rickettsia prowazekii*. While most screens will be performed within the same genome from which the known sites are generated, it is possible to search other genomes using the same matrix. However, the benefits of searches of this nature still remain unexplored. Following the submission of a search query, GRASP-DNA will notify the researcher of the search completion via email referencing a temporary URL that will house the results and persist for approximately two weeks. The results can be viewed and sorted according to ranking or location within the genome. For every site the location and orientation of the two closest open reading frames and their translational start sites are also provided. The open reading frames can be listed by their locus identification numbers or by their genetic annotations for most genomes. (As an aside note the program can also be used to identify the location of all restriction sites for a given restriction endonuclease by entering in the cleavage site sequence and searching the genome of choice. All of the exact matches to the cleavage site and their precise locations will appear at the top of the results list.)

Example Screen for the FruR DNA-Binding Protein

To illustrate an application of GRASP-DNA we present a screen of the *E. coli* genome for potential binding sites of FruR, the fructose repressor. FruR is a global transcriptional regulatory protein involved in the control of carbon and energy metabolism (Saier and Ramseier, 1996). Eleven experimentally confirmed binding sites are known for FruR (Ramseier *et al.*, 1995), which specifically modulate the activation of genes involved in oxidative and gluconeogenic carbon flow and the repression of genes involved in fermentative carbon metabolism. The 14 base pair sequence of all eleven binding sites was entered into the GRASP-DNA interface along with identifiers for each of the sites, and was used to construct a position weight matrix (Figure 1). GRASP-DNA then uses this matrix to score every possible 14 base pair sequence window of the *E. coli* genome on both the leading and lagging strand. A sample of the top scoring sites generated from the screen is provided in Figure 2.

All eleven sites scored in the top 50 of the 9,278,442 possible sequence windows in the genome. Any sequence site that matches an input sequence used to generate the position weight matrix is highlighted in yellow to rapidly identify the known binding sites. While it is expected that these sites used to generate the matrix should appear at the top of the list, there are a number of additional sites that may have functional importance based on the location of the binding site relative to the transcriptional and translational start sites of nearby genes. An example of this would be the site ranked #4 which may indicate the presence of a third regulatory binding site for the fruB operon in *E. coli*.

In a retrospective analysis the six binding sites for

FruR, which were published 2 years before the discovery of 5 additional sites (Ramseier *et al.*, 1993), were used to screen the *E. coli* genome. Using either information theory or statistical mechanical methods with the default parameter setting, GRASP-DNA screened the *E. coli* genome and identified the other 5 binding sites within the top 300 sites (see Table 1). Thus the previously known sites could have been used to assist in identifying the other sites more readily with the assistance of a weight matrix based computational analysis and access to the whole genome.

Conclusion

In general the results from GRASP-DNA can be used to help point in the direction of putative binding sites that may affect genetic regulation and assist in further identifying genetic regulatory networks. These predictions can then be further explored using genome-scale techniques such as cDNA microarrays or oligonucleotide arrays to assess transcriptional activity of genes under the appropriate experimental and control conditions (Roth *et al.*, 1998). The complementary application of computational algorithms to find regulatory binding sites and experimental transcriptome analysis has been recently highlighted (Bucher, 1999). Ultimately, predictions made from computational analyses of transcriptional binding sites must be validated with biochemical assays to determine the precise nature of the protein-DNA binding. With the assistance of web-accessible computational tools such as GRASP-DNA and recent experimental technologies, it is now possible to make educated decisions on the likelihood of a binding site existing before the time and effort is spent to characterize the precise binding interactions in vitro. We look forward to the use of GRASP-DNA for various purposes, and welcome further comments and suggestions to assist in the development of this application in the future.

References

- Berg, O.G. and Von Hippel, P.H. 1987. Selection of DNA binding sites by regulatory proteins. *J. Mol. Biol.* 193: 723-750.
- Bucher, P. 1999. Regulatory elements and expression profiles. *Current Opinion in Structural Biology* 9: 400-407.
- Claverie, J.M. 1994. Some useful statistical properties of position-weight matrices. *Comput. Chem.* 18: 287-294.
- Claverie, J.M., and Audic, S. 1996. The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.* 12: 431-439.
- Frech, K., Quandt, K., and Werner, T. 1997. Finding protein-binding sites in DNA sequences: the next generation. *Trends Biochem. Sci.* 22: 103-104.
- Frech, K., Quandt, K., and Werner, T. 1997. Software for the analysis of DNA sequence elements of transcription. *Comp. Appl. Biosci.* 13: 89-97.
- Heinemeyer, T., Chen, X., Karas, H., Kel, A.E., Kel, O.V., Liebich, I., Meinhardt, T., Reuter, I., Schacherer, F., and Wingender, E. 1999. Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucl. Acids Res.* 27: 318-22.
- Lewis, L.K. 1994. Identification of high affinity binding sites for LexA which define new DNA damage-inducible genes in *Escherichia coli*. *J. Mol. Biol.* 241: 507-523.
- Magasanik, B. 1993. The regulation of nitrogen utilization in enteric bacteria. *J. Cell. Biochem.* 51: 34-40.
- Ramseier, T.M., Bledig, S., Michotey, V., Feghali, R., and Saier, M.H. 1995. The global regulatory protein FruR modulates the direction of carbon flow in *Escherichia coli*. *Mol. Microbiol.* 16: 1157-1169.
- Ramseier, T.M., Nègre, D., Cortay, J. C., Scarabel, M., Cozzone, A. J., and Saier, M. H. 1993. In vitro binding of the pleiotropic transcriptional regulatory protein, FruR, to the fru, pps, ace, pts and icd operons of *Escherichia coli* and *Salmonella typhimurium*. *Journal of Molecular Biology* 234: 28-44.
- Robison, K., McGuire, A.M., and Church, G.M. 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete

- escherichia coli K-12 genome. *J. Mol. Biol.* 284: 241-254.
- Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotech.* 16: 939-945.
- Saier, M.H. 1989. Protein phosphorylation and allosteric control of inducer exclusion and catabolite repression by the bacterial phosphoenolpyruvate:sugar phosphotransferase system. *Microbiol. Rev.* 53: 109-120.
- Saier, M.H., and Ramseier, T.M. 1996. The catabolite repressor/activator (Cra) protein of enteric bacteria. *J. Bacteriol.* 178: 3411-3417.
- Salgado, H., Santos, A., Garza-Ramos, U., Van Helden, J., Díaz, E., and Collado-Vides, J. 1999. RegulonDB (version 2.0): a database on transcriptional regulation in *Escherichia coli*. *Nucl. Acids Res.* 27: 59-60.
- Schneider, T.D. 1997. Information content of individual genetic sequences. *Journal of Theoretical Biology* 189: 427-41.
- Stormo, G. 1990. Consensus patterns in DNA. *Methods Enzymol.* 183: 211-221.
- Stormo, G.D. 1988. Computer methods for analyzing sequence recognition of nucleic acids. *Ann. Rev. Biophys. Biophys. Chem.* 17: 241-263.
- Thieffry, D., Salgado, H., Huerta, A.M., and Collado-Vides, J. 1998. Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12. *Bioinformatics* 14: 391-400.
- Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M., and Pontoglio, M. 1997. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.* 266: 231-245.
- Wanner, B.L. 1993. Gene regulation by phosphate in enteric bacteria. *J. Cell. Biochem.* 51: 47-54.
- Wasserman, W.W. and Fickett, J.W. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* 278: 167-181.