Nucleic Acids Research, 1998, Vol. 26, No. 1 55–59

RegulonDB: a database on transcriptional regulation in *Escherichia coli*

Araceli M. Huerta, Heladia Salgado, Denis Thieffry and Julio Collado-Vides*

Centro de Investigación sobre Fijación de Nitrógeno, UNAM A.P. 565-A Cuernavaca, Morelos 62100, México

Received September 15, 1997; Revised and Accepted October 15, 1997

ABSTRACT

RegulonDB is a DataBase that integrates biological knowledge of the mechanisms that regulate the transcription initiation in *Escherichia coli*, as well as knowledge on the organization of the genes and regulatory signals into operons in the chromosome. The operon is the basic structure used in RegulonDB to describe the elements and properties of transcriptional regulation. The current version contains information around some 500 regulation mechanisms, essentially for sigma 70 promoters.

INTRODUCTION

The construction of data banks to gather information about biological processes and structures is a growing and active area of research as testified by several specialized databases such as EcoCyc and MetalgenDB, two databases on metabolic pathways in *Escherichia coli* (1,2); ECO2DBASE, a 2-D gel database of *E.coli* (3); and TRANSFAC, a database on transcription factors in eukaryotic cells (4). In this paper, we present RegulonDB, a database on transcriptional regulation in *E.coli*. This database contains information on regulatory features such as promoters, binding sites for regulatory proteins, as well as their associated regulated genes organized into operons, and regulons. Although partial information exists about promoter sequences (5) and regulatory mechanisms (6), this information has not been integrated into a computer-accessible database. Among other properties, the database contains information on the relative position of regulatory sites, the transcription initiation, the distance to the beginning of the transcribed gene, and their coordinates within the completed *E.coli* genome (7). Regulatory interactions are associated with the experimental evidence that supports them and the literature source. The structure of the database together with a description of the main biological properties of gene regulation contained in the database are described in the next sections.

METHOD

RegulonDB was developed using a *relational DataBase scheme* (8). A commercial software for building relational databases in Macintosh platform, 4thDimension (4thD) was used (9). Scripts were written into 4thD to generate the graphical interface for the user, and the interface for updating the data. In addition, Reference Manager (10) was used to organize a parallel literature database. The references included into RegulonDB were initially loaded into Reference Manager, exported from there in a homogenous format, and loaded into RegulonDB. Information was gathered from different sources, such as reviews on gene regulation (6), on promoter sequences (5), other databases (11), searches in Medline and, in a smaller fraction, from GenBank (12). Scripts were written in Perl (13) to manipulate this information.

THE STRUCTURE OF THE DATABASE

Operons and regulatory interactions

RegulonDB essentially describes the interactions of regulatory proteins and their associated binding sites, as well as the organization of regulatory features (sites, genes, and promoters) into their associated operons and regulons. The definition of operon that this database is built upon is that of a polycistronic transcribed unit with its associated regulatory sites, whereas a regulon is defined as a group of operons controlled by one regulator (14–16). As discussed below, operons and their internal structure are described by tables in a hierarchical organization. On the other hand, a regulatory interaction (RI) can be defined as a quadrutuple RI(P, RP, S, F) where P is the regulated promoter, RP, the regulatory protein, S, the site where the regulator binds, and F, the function or regulatory effect on the regulated promoter. The table called **regulatory interactions** is the core for this description into the relational model as described below. (To facilitate comprehension, tables will be indicated in boldface and fields in italics in the text.)

The relational structure

A relational DB system consists of a set of a tables and a set of relations between tables, where each table contains codified information (attributes) about a particular object, such as operons, promoters, regulatory interactions, etc.; and a relation between two objects is defined by listing the two objects as an ordered pair. The relational design of RegulonDB involved modeling of the internal structure of operons into regulatory sites, promoters and genes, on the one hand; of the physical interactions of the molecules involved in transcription regulation, on the other hand. This model is described in Figure 1.

* To whom correspondence should be addressed. Tel: +52 7 313 2063; Fax: +52 7 317 5582; Email: ecoli-reg@cifn.unam.mx

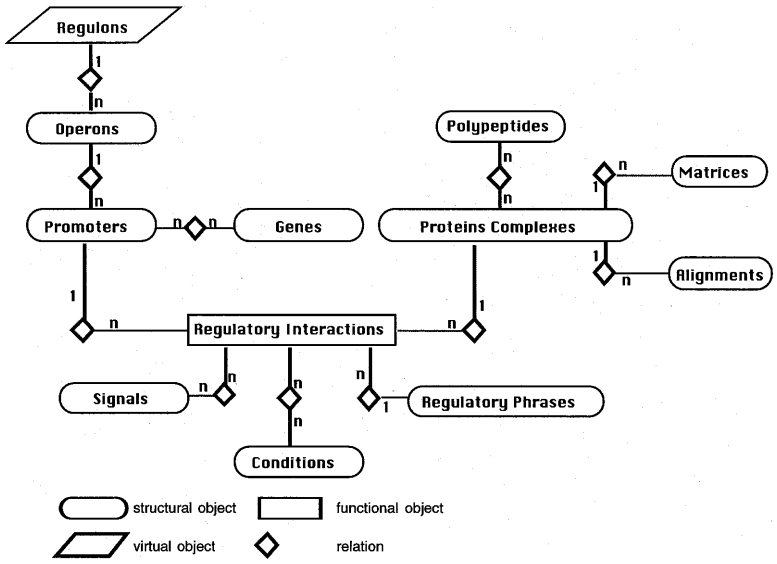

Figure 1. The conceptual model of RegulonDB. This model distinguishes virtual, functional and structural objects; one-to-n and n-to-n relationships between the objects. See text for more description of the model.

Figure 1 shows two types of relationships, simple and complex ones. A simple relationship is a one-to-n relationship between two objects, for example, the relationship between **operons** and **promoters**; certainly, one operon can contain several promoters, and one promoter belongs to a single operon. Complex relationships between two objects need the creation of an intermediate table that does not correspond to any object; these are n-to-n relationships (see the diamonds in Fig. 1). An example of one complex relationship is presented below. As a first approximation of the model of gene regulation underlying the database, we can say that it contains three types of objects: those that define 'apparent physical entities' or biological structures (i.e. **operons**, **promoters**, **genes**, **signals**, **polypetides**, **protein complexes**, **matrices** and **alignments**); the functional table that describes the complex **regulatory interactions**, and one virtual object.

We call the first type 'apparent physical entities' because in a deeper consideration we see it is more adequate to think of operons, promoters and even genes, as functional properties mapped on specific sequences, more than DNA sequences with inherent properties (for a more detailed discussion see 17, and a related discussion in 18). The table of **regulatory interactions** models the quadru-tuple RI function. Finally, **regulons** are virtual objects since there is no regulon table in the database, they are generated with a program using information contained in other tables.

The diagram of the implementation of this model with 4thD is shown in Figure 2, where tables with their attributes, and relationships among tables, are described. A one-to-n relationship is described by an arrow between two tables in Figure 2, and complex relationships are modeled by the so-called 'link-tables'.

One example of a complex relationship is the one among genes and promoters described by the **genes-link** table. Certainly, one gene can be transcribed from several promoters, and one promoter can initiate transcription of several genes. This table permits to model in the database an operon such as the one shown in Figure 3, where promoter A (pA) transcribes genes G1, G2 and G3, and promoter B (pB) transcribes G2 and G3. Each promoter is described separately within the table of **promoters**, and each gene is described separately within the table of genes. The **genes-link** table contains the relationships described by the pairs (pA, G1), (pA, G2), (pA, G3) and (pB, G2), (pB, G3). Additional link tables are for instance the **signals-link** table that connects signaling metabolites with the regulatory interactions, and the **conditions-link** table connecting physiological conditions with their associated regulatory interactions.

It is important to note that these simple and complex interactions are in fact defining a hierarchy within the tables in the model. One example is the hierarchical relation between **regulatory interactions** on one hand, and **promoters** and **proteins** on the other. A regulatory interaction cannot be described if a regulated promoter and its associated regulatory protein have not been previously described. Similarly, **genes** and **promoters** must belong to a given operon, and therefore, only genes and promoters which can be defined as part of an operon are included in the database. This hierarchical organization is reflected in practice in our programs for updating of information in the database. Examples of the biological complexity of operon organization and regulatory interactions are presented in more detail in a subsequent paper.

TABLES AND THEIR RELATIONS

In terms of the biology, RegulonDB comprises five main tables: **regulatory interactions**, **protein complexes**, **promoters**, **genes** and **operons**. All of them have a field *external_database* with information of references to other DataBase.

The table of **regulatory interactions** contains its identifier *regulatory_interaction_id*, the *ID*s (identifiers) corresponding to the two objects involved in a regulatory interaction: the regulatory protein (*protein_complex_id*), and the promoter that is being regulated (*promoter_id*); the information on the site is part of the table itself (*site_id*), this number groups all regulatory interactions that share the same physical site in the DNA, all these interactions will have the same site_id. Other attributes are

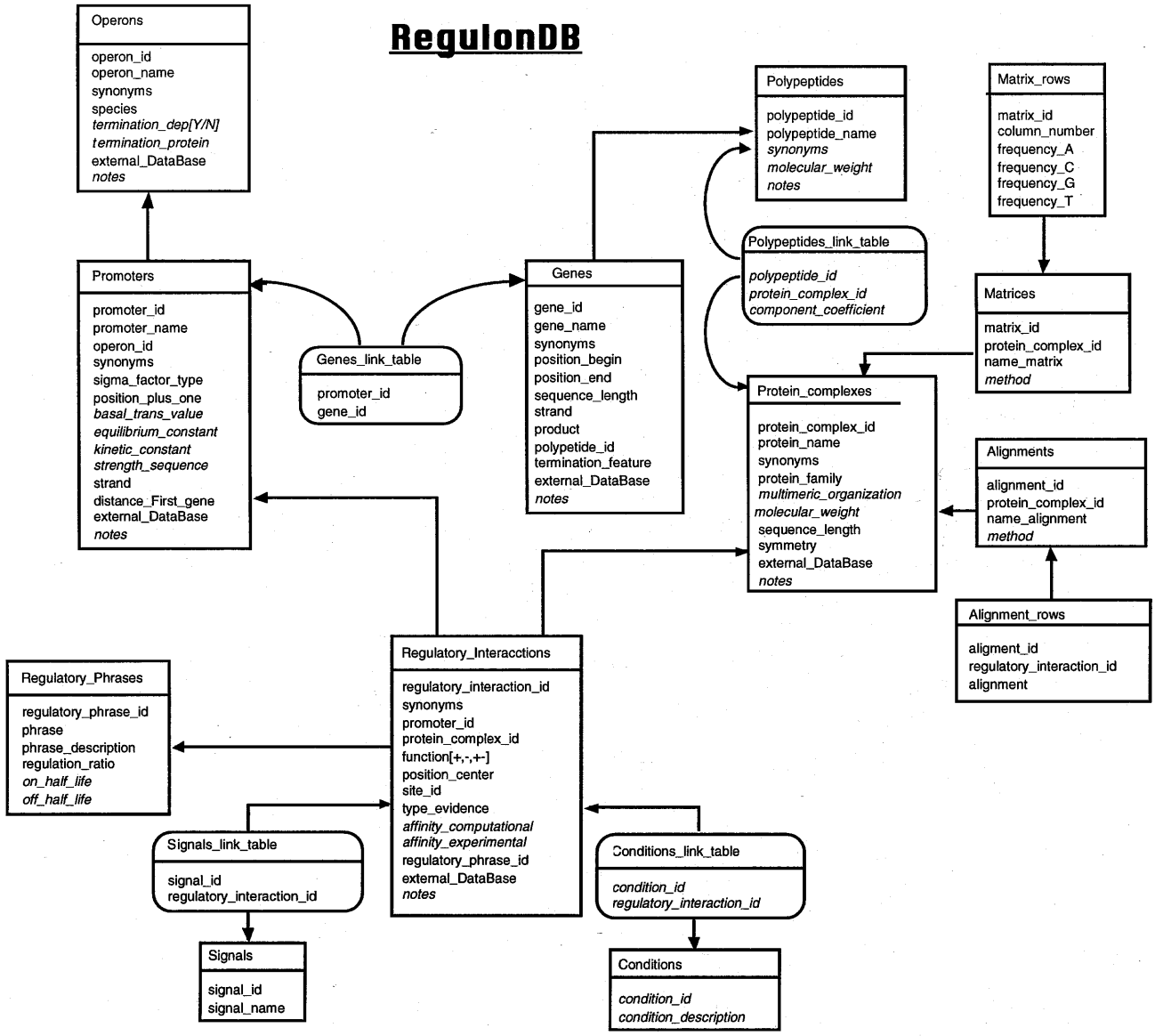

Figure 2. The structure of RegulonDB. The conceptual model of Figure 1 as shown in 4thD. Link-tables are represented with round corners, the others are represented with sharp corners.

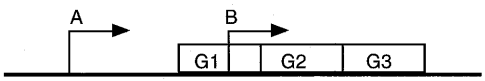

Figure 3. A complex operon with more than one promoter. A scheme of the a operon with two promoters and three genes.

synonyms, *position_center* from transcription initiation, as well as the *function* (activator, repressor or dual). Another important feature in this table is the *type_evidence* associated to a given regulatory interaction. These are organized into four classes following (6): whether there is (i) mutational experiments proving the protein–DNA binding site interaction, (ii) specific binding of the purified protein, (iii) evidence of binding with non-purified protein, or (iv) simple sequence similarity with other sites for a regulatory protein. Additional features not yet filled include *affinity* between the protein and the site (both experimentally and computationally derived), *regulation_ratio* (between the level of basal and regulated expression). A set of regulatory interaction can form a regulatory phrase (see the **regulatory phrase** table below for more explanation), the *regulatory_phrase_id* identifies this phrase. Another attribute is that of *notes* containing comments about the regulatory interaction.

The table of **protein complexes** contains the *protein_complex_id*, the *protein_name*, *synonyms*, and properties of the regulatory proteins such as their evolutionary family membership in *protein_family*, the *molecular_weight*, the *sequence_length*, *symmetry* (direct repeat, inverted repeat or asymmetric) and their *multimeric_organization*, a number indicating if the protein is a

monomer, dimer, tetramer, etc. *Notes* contains comments about the protein complex.

The table of **promoters** contains the identifier *promoter_id*, the *promoter_name*, *synonyms*, and the *operon_id* to which it belongs, as well as information on *strand* (reverse or forward) orientation, the *sigma_factor_type*, the position of transcription initiation in the genome (*position_plus_one*), and the *distance_first_gene* that is transcribed within the operon. Additional features not yet completed include *basal_trans_value* that is a number indicating the rate of transcription of the promoter in the absence of any regulation; the *equilibrium_constant* for the binding of the RNA polymerase to the promoter; the *kinetic_constant* of the transition from closed to open complex, and *strength_sequence*, the score of the promoter using the weight matrix for the given type of promoter. *Notes* contains comments about the promoter.

The table of **genes** contains the *gene_id*, *gene_name*; and *synonyms*, as well as properties describing the position of genes in the genome (*initial_position* and *end_position*), the size of the gene *sequence_length*, *strand*, *product*, and *termination_features* (if the gene contains a termination signal). *Polypeptide_id* links the protein complex with the polyptides that form it. We have not collected much information on the regulated genes and the functional properties of their *products* since that information can be found in other databases such as SWISSPROT or EcoCyc. We plan to include EcoCyc accession numbers to all genes. *Notes* contains comments about the gene.

The table of **operons** contains the *operon_id*, *operon_name*, *species*, and *synonyms*. Additional features not yet completed include *termination_dep* (termination depended; one of, yes or not) and *termination_protein* (name of terminator protein). *Notes* contains comments about the operon.

Additional tables that enrich the description on gene regulation are: **regulatory phrases**, **conditions**, **signals**, **polypeptides**, **matrices**, **matrix_rows**, **alignments**, and **alignment_rows**; altogether with the link tables: **genes_link_table**, **signals_link_table**, **conditions_link_table**.

When a single promoter is regulated by several regulatory interactions, these can be grouped, since they participate in a single regulatory mechanism or phrase affecting coordinately the promoter activity. These phrases are described in the **regulatory phrases** table in the database; regulatory phrases are classified (*phrase_description*), as homologous or heterologous; positive, negative, or dual. A group of regulatory interactions is homologous if all sites bind the same protein and heterologous otherwise. A group is negative if all sites repress the promoter, positive if all activate the promoter, or dual if it contains activator and repressor sites. Other fields are *phrase_ID*, the *phrase* (a description of the set of sites that define the phrase) , the *regulation_ratio* (number of times the given promoter is turned on or off), *on_half_life (*half life to turn on the regulated gene), and *off_half_life* (half life to turn off the regulated gene). The three last are not yet filled in.

The database encodes in the **conditions** table the physiological conditions under which mechanisms are turned on or off in the cell, the fields are: *condition_id*, and *condition_description*. This table is not completed yet. The signal metabolite that is responsible for the relevant change of conformation of the regulatory protein (i.e. allolactose for LacI, cAMP for CRP, etc.) is described in the **signals** table, the fields in this table are *signal_id*, and *signal_name*. Links between these features and the regulatory interactions are located within the **signals_link_table** and **conditions_link_table** tables using the *regulatory_interaction_id* and *conditions_id* or *signals_id* in each case. The genes-link_table lists, for a given promoter, the *ids* of the genes that are transcribed from that promoter (*promoter_id, gene_id*). The tables for **matrix_rows** (*matrix_id*, *column_number*, *frequency_in_A*, *frequency_in_G*, *frequency_in_C* and *frequency_in_T*) and **alignment_rows** (*alignment_id*, *regulatory_interaction_id* and *alignment_row*) describe the weight matrix and the associated multiple alignment for a set of binding sites (when there are enough of them) for a given protein. The matrices were generated using the program *Wconsensus* (19). The matrix selected is the one with the lowest expected frequency that includes all the known functional sites for a protein. The **matrices** and **alignments** tables connects this information to the Data Base using the *protein_complex_id*, *matrix_id* or *alignment_id* and the *method* used to generate the matrix and the alignment. The table for **polypetides** contains *polypeptide_id*, *polypeptide_name*, *synonyms*, *notes* and the *molecular_weight*. The **polypeptides-link_table** links the polypetides with the protein complexes and contains the *component_coefficient* (the number the monomers that contribute to the protein complex), the id's *polypeptide_id*, and *protein_complex_id*.

As mentioned before not all these fields are comprehensively completed. Figure 2 contains in plain text those fields for which there is information in some cases, and in italics those for which no information is present within the current version 1.0 of RegulonDB. A comprehensive description of each field of the database can be found in the documentation on our ftp site. In Table 1, a summary of the amount currently contained in the main fields of the database is presented.

Note that though several regulatory interactions do not have an explicit reference associated, at least one reference to an external literature database such as a Medline accession number, or a GenBank accession number is associated to each regulatory interaction. In this way, we made sure that every regulatory interaction in RegulonDB is supported by at least one reference to the literature.

Table 1. Information contained in RegulonDB release 1.0

Object	Number
Regulons	99
Regulatory interactions	533
Polypeptides	192
Protein complexes	99
Genes	542
Operons	292
Promoters	300
Ext_DB_References[a]	2050
Authors	298
Signals	35

[a]Basically to Medline and GenBank.

HOW TO ACCESS REGULON DB

The database can be obtained in order to be installed to run on a Macintosh directly from the web at http://www.cifn.unam.mx/Computational_Biology/regulondb . Compressed files available in either Binhex or MacBinary II format can be obtained. A document describing in detail the tables and fields of the database is also available. This same web page will permit direct access to the database. It can also be obtained by anonymous ftp from: ftp.cifn.unam.mx/pub/software/mac. There are three files: RegulonDB.readme with information on its installation, RegulonDB.sea and RegulonDB.ps. Once the files are obtained, they can be downloaded, decompressed, and installed on a Macintosh with at least 8 MB in RAM and 9 MB free in hard disk.

ACKNOWLEDGEMENTS

J.C.-V. acknowledges Temple Smith for being invited to the BioMolecular Engineering Research Center, Boston University, where in collaboration with Kathleen Klose, the first design of this database was made. We acknowledge Peter Karp for discussions during the design of the database, Ernesto Pérez-Rueda for information on protein families, and Victor Del Moral for computer support. This work was supported by grants from DGAPA-UNAM and Conacyt to J.C.-V.

REFERENCES

1 Karp,P., Riley,M., Paley,S.M. and Pelligrini-Toole,A. (1996) *Nucleic Acids Res.*, **24**, 32–39 [see also this issue (1998) *Nucleic Acids Res.* **26**, 50–53].
2 Rouxel,T., Danchin,A. and Henaut,A. (1993) *Nucleic Acids Res.*, **9**, 315–324.
3 VanBogelen,R.A., Abshire,K.Z., Pertsemlidis,A., Clark,R.L. and Neidhardt,F.C. (1996) In Neidhardt,F., Curtiss,R.I., Gross,C.A., Ingraham,J.L. and Riley,M. (eds), *Escherichia coli and Salmonella typhimurium Cellular and Molecular Biology.* American Society for Microbiology, Washington D.C., Vol I, pp. 2067–2117.
4 Wingender,E., Dietze,P., Karas,H. and Knueppel,R. (1996) *Nucleic Acids Res.*, **24**, 238–241 [see also this issue (1998) *Nucleic Acids Res.* **26**, 362–367].
5 Lisser,S. and Margalit,H. (1993). *Nucleic Acids Res.*, **21**, 1507–1516.
6 Collado-Vides,J., Magasanik,B. and Gralla,J.D. (1991) *Microbiol. Rev.*, **55**, 371–394.
7 Blattner,F.R., Plunkett III,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G., *et al.* (1997) *Science*, **277**, 1453–1462.
8 Date,C.J. (1990) *An Introduction to Database Systems, Volume I.* Addison-Wesley Publishing Company, Inc., Reading, Massachusetts.
9 *4th Dimension* is a registered trademark of ACI/ACI US, Inc. 20883 Stevens Creek Blvd., Cupertino, CA 95014.
10 *Reference Manager* is a copyright of Research Information System, Inc. Camino Corporate
11 DPInteract, Robinson,K., and Church,G.M., http://arep.med.harvard.edu/dpinterac/
12 Benson,D.A., Bogusky,M., Lipman,D.J. and Ostell,J. (1996) *Nucleic Acids Res.*, **24**, 1–5 [see also this issue (1998) *Nucleic Acids Res.* **26**, 1–7].
13 Wall,L. and Schwartz,R.L. (1991) *Programming Perl.* O'Reilly and Associates, Inc., Sebastopol, Ca.
14 Beckwith,J. (1996) In Neidhardt,F., Curtiss,R.I., Gross,C.A., Ingraham,J.L. and Riley,M. (edS), *Escherichia coli and Salmonella typhimurium. Cellular and Molecular Biology.* American Society for Microbiology, Washington D. C., Vol. I, pp. 1227–1231.
15 Jacob,F. and Monod,J. (1961) *J. Mol. Biol.*, **3**, 318–356.
16 Neidhardt,F. (1987) In Neidhardt,F., Curtiss,R.I., Gross,C.A., Ingraham,J.L. and Riley,M. (eds), *Escherichia coli and Salmonella typhimurium. Cellular and Molecular Biology.* American Society for Microbiology, Washington D. C., Vol. I, pp. 1313–1317.
17 Collado-Vides,J., Huerta,A.M. and Klose,K. (1997) In Hofestadt,R., Lengauer,T., Löffler,M. and Schomburg,D. (eds), *Bioinformatics - Lecture Notes in Computer Science,* Vol. 1278, Springer-Verlag, Berlin, pp. 72–78.
18 Robbins,R. (1996) In Collado-Vides,J., Magasanik,B. and Smith,T. (eds), *Integrative Approaches to Molecular Biology.* MIT Press, Cambridge, MA., pp. 91–112.
19 Hertz,G.H. and Stormo,G.D. (1995) In Lim,H.A. and Cantor,C.R. (eds), *Proceedings of the 3rd International Conference on Bioinformatics and Genome Research.* World Scientific Publications, Singapore, pp. 199–214.