Université d'Ottawa · University of Ottawa

PERMISSION DE REPRODUIRE ET DE DISTRIBUER LA THÈSE

PERMISSION TO REPRODUCE AND DISTRIBUTE THE THESIS

NOM DE L'AUTEUR / NAME OF AUTHOR:	Robert Gerald Beiko
ADRESSE POSTALE / MAILING ADDRESS:	1812-1541 Riverside Drive #9/31 Glen Rd. Ottawa, Ontario K1G 4E2 Toowong, QLD. Australia 4066
GRADE / DEGREE:	**ANNÉE D'OBTENTION / YEAR GRANTED**
Ph. D- Biology	2003
TITRE DE LA THÈSE / TITLE OF THESIS:	
Evolutionary Computing Strategies for the Detection of Conserved Patterns in Genomic DNA	

L'auteur permet, par la présente, la consultation et le prêt de cette thèse en conformité avec les règlements établis par le bibliothécaire en chef de l'Université d'Ottawa. L'auteur autorise aussi l'Université d'Ottawa, ses successeurs et cessionnaires, à reproduire cet exemplaire par photograhie ou photocopie pour fins de prêt ou de vente au prix coûtant aux bibliothèques ou aux chercheurs qui en feront la demande.

Les droits de publication par tout autre moyen et pour vente au public demeureront la propriété de l'auteur de la thèse sous réserve des règlements de l'Université d'Ottawa en matière de publication de thèses.

The author hereby permits the consultation and the lending of this thesis pursuant to the regulations established by the Chief Librarian of the University of Ottawa. The author also authorizes the University of Ottawa, its successors and assignees, to make reproductions of this copy by photographic means or by photocopying and to lend or sell such reproductions at cost to libraries and to scholars requesting them.

The right to publish the thesis by other means and to sell it to the public is reserved to the author, subject to the regulations of the University of Ottawa governing the publication of theses.

N.B. LE MASCULIN COMPREND ÉGALEMENT LE FÉMININ

Aug 19, 2003 — Robert G Beiko

DATE (AUTEUR) SIGNATURE (AUTHOR)

Université d'Ottawa · University of Ottawa

Université d'Ottawa · University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

BEIKO, Robert

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

Ph. (Biology)

GRADE - DEGREE

Biology

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Evolutionary Computing Strategies for the Detection of Conserved Patterns in Genomic DNA

R. Charlebois

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

S. Findlay

D. Hickey

I. Lambert

A. Roger

M. Turco

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Evolutionary Computing Strategies for the Detection of Conserved Patterns in Genomic DNA

Robert G. Beiko

Thesis submitted to the
School of Graduate Studies and Research
University of Ottawa
In partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Ottawa-Carleton Institute of Biology

Abstract

The detection of regulatory sequences in DNA is a challenging problem, especially when considered in the context of whole genomes. The degree of sequence conservation of regulatory protein binding sites is often weak, and the sites are obscured by surrounding intergenic sequence. Since structural interactions are vital for protein-DNA interactions, structural representations of regulatory sites can yield a more accurate model and a better understanding of within-site variability. However, the use of multiple alternative representations of DNA introduces a requirement for novel algorithms that can create and test different combinations of DNA features.

The Genetic Algorithm Neural Network (GANN) was designed to identify combinations of patterns that can be used to distinguish between different classes of training sequence. GANN trains a set of artificial neural networks to classify sets of sequence using either backpropagation or a genetic algorithm, and uses an 'outer genetic algorithm' to choose the best inputs from a pool of DNA features that can include sequence, structure, and weight matrix representations. When trained with a subset of upstream sequences from a whole genome, GANN was able to detect patterns such as the Shine-Dalgarno sequence in *Escherichia coli* K12, and sequences consistent with archaeal promoters in the archaeon *Sulfolobus solfataricus* P2.

The Motif Genetic Algorithm (MGA) constructs motif representations by concatenating minimal units of DNA sequence and structure. This algorithm was used to model conserved patterns in DNA, including the binding sites for *E. coli* cyclic AMP activated protein (CAP), integration host factor (IHF), and two different promoter types recognized by alternative bacterial sigma factors. The CAP models were used to detect other putative binding sites in upstream regions of the *E. coli* K12 genome, while attempts to train an accurate model of IHF binding sites revealed an important role for structural representations in motif modeling.

Résumé

La détection des éléments de régulation en ADN est un problème important dans le contexte des génomes entiers. Le degré de conservation des accepteurs de protéines régulatrices est souvent faible, et les emplacements sont obscurcis par les séquences intergéniques adjacentes. Puisque les interactions structurales sont essentielles pour des interactions protéine-ADN, les représentations structurales des éléments de régulation peuvent rapporter un modèle plus précis et une meilleure compréhension de la variabilité entre les éléments qui sont reconnus par la même protéine. Cependant, l'utilisation des représentations alternatives multiples de l'ADN présente une condition pour des algorithmes nouveaux qui peuvent créer et examiner différentes combinaisons des représentations de l'ADN.

Le réseau neurologique d'algorithme génétique (GANN) a été conçu pour identifier des combinaisons des modèles qui peuvent être employés pour distinguer différentes classes de séquences. GANN forme un ensemble de réseaux neurologiques artificiels pour classifier des ensembles de séquence en utilisant le backpropagation ou un algorithme génétique, et emploie un algorithme génétique 'extérieur' pour choisir les meilleures entrées d'une groupe des entrées d'ADN qui peuvent inclure l'ordre, la structure, et les représentations de matrice. Une fois entrainé avec un sous-ensemble de séquences en avant des gènes d'un génome entier, GANN pouvait détecter des modèles tels que le séquence Shine-Dalgarno dans *Escherichia coli* K12, et les séquences conformés aux promoteurs archaeals dans l'espèce *Sulfolobus solfataricus* P2.

L'algorithme génétique de motif (MGA) construit des représentations de motif en enchaînant les unités minimales des séquences et de la structure de l'ADN. Cet algorithme a été employé pour représenter les modèles conservés en ADN, y compris les accepteurs pour la protéine activée par adénine cyclique de *E. coli* (CAP), 'Integration Host Factor' (IHF), et deux types différents de promoteur reconnus par des facteurs sigma bactériens alternatifs. Les modèles de CAP ont été employés pour détecter d'autres accepteurs putatifs dans des régions en avant des gènes du génome de *E. coli* K12, alors que les tentatives de former un modèle précis des accepteurs d'IHF indiquaient un rôle

important pour les représentations structurales.

Acknowledgments

I am grateful for the contributions of many people to this work. First and foremost, my supervisor Robert L. Charlebois has been an excellent teacher and mentor, whose ability to think in unconventional ways has yielded an interesting and productive four and a half years. The members of my supervisory committee, Donal Hickey, Iain Lambert, and Alain St-Amant, have contributed useful questions and suggestions (which are hopefully addressed somewhere in this document). I would also like to thank Scott Findlay and Claude Desruisseaux for their help in dealing with gigabytes of data and DNA structure, respectively.

Others who have helped out along the way include Guy Drouin, Cristofre Martin, Julie Chapados, Stephen Stockton, Anthony Francis and the members of the Bioinformatics Discussion Group. The other two members of the 'Three Amigos', Greg Singer and Mehrdad Hajibabaei, have been great friends and excellent colleagues with the indispensable ability to find good food at reasonable prices.

The effort required to finish this thesis would have been far greater without the help and support of my wife, Jennifer Georgeff. And finally, I would like to thank my parents Jerry and Kitty for their support and their commitment to my education from day one. This effort is dedicated to them.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural Network
bp	base pair or base pairs
CAP	Cyclic AMP activated protein
DFA	Discriminant Function Analysis
GANN	Genetic Algorithm Neural Network
IGA	Inner Genetic Algorithm
IHF	Integration Host Factor
MGA	Motif Genetic Algorithm
MLP	Multi-Layer Perceptron
nt	nucleotide or nucleotides
OGA	Outer Genetic Algorithm
ORF	Open Reading Frame
RNAP	RNA Polymerase

[1]

Introduction

The Regulation of Gene Expression

The set of proteins and RNA molecules required by an organism are encoded by genes in its genome. Gene expression is the process by which these genes are converted to functional molecules within the cell, first by transcription from DNA to RNA, then, if the gene encodes a protein, by translation from RNA to protein. While accessory factors are often necessary, the vital processes of transcription and translation are each dependent on one critical component. Transcription is dependent on RNA polymerase, a multi-subunit protein that initiates transcription by melting the upstream DNA, and uses the single-stranded DNA as a template to assemble a molecule of messenger RNA (mRNA). Translation requires the ribosome, a complex of ribosomal proteins and RNAs that identifies and binds either to a short, purine-rich sequence in the mRNA near the translation start codon or to a set of proteins that interact with a modified ribonucleotide at the 5′ end of the mRNA (reviewed in Kozak, Gene 1999), and recruits a series of amino acids to create a protein molecule.

Some gene products need not be as abundant within the cell as others, and some are only needed under specific physiological conditions. While underexpression of a gene yields a shortfall of the required product, overexpression of a gene leads to a waste of resources and energy, and can also interfere with important cellular processes. To control the level of gene expression, cells have developed a number of regulatory processes that determine under what conditions and how often a gene product should be assembled.

These regulatory processes can be exerted at every step of gene expression. A primary condition of activation is that the region of DNA containing a gene must be accessible by the RNA polymerase and any other required factors. A gene can thus be sequestered away from the transcription machinery, which occurs when a gene is located

in the compacted center of the nucleoid in bacteria (Reviewed in Ishihama *et al.*, 1998) and when a gene is silenced by a tightly-packed chromatin structure in some Archaea and Eukarya (Kadam and Emerson, 2002; Horn and Peterson, 2002). Access can also be impaired by proteins bound at or near the promoter, where RNA polymerase must bind the DNA to initiate transcription. Repressor proteins are frequently used in Bacteria and Archaea, and inhibit transcription by binding either the promoter region itself, or a stretch of DNA downstream from the promoter (Müller-Hill 1998). Repressor proteins are also found in eukaryotes, but inactivation of transcription is more commonly achieved with a series of histones bound to the upstream DNA as well as the sequence to be transcribed (Kadam and Emerson, 2002). The use of such methods to reduce or silence transcription is termed *negative control.*

Different levels of gene expression are also achieved through variable affinity between the regulatory protein and its target binding sequence. A single regulatory molecule or polymerase unit can typically interact with any of a number of related target DNA sequences, with higher association rates for some target sequences than for others. This variability allows the same regulatory protein or polymerase to exert different degrees of control over different transcription units, yielding a set of co-expressed operons called a *regulon* (Neidhardt and Savageau, 1996; Hecker and Volker, 1998).

With *positive control,* activator proteins are used to enable transcription or increase the rate at which it occurs. These proteins are frequently used in Archaea and Bacteria, where a weak interaction between RNA polymerase and the promoter can be stabilized through protein-protein interactions between RNA polymerase and one or more activator proteins (Rhodius and Busby, 1998). This mode of control predominates in eukaryotes, where in addition to bacterial-like activators, another class of proteins that bind enhancer regions are often required to deactivate and remove the histones that are bound near the promoter (Lee and Young, 2000).

Regulation of gene expression also occurs at other steps in the conversion of a signal from DNA to mature RNA or protein. Translational regulation can be exerted either in a general way, such as through the limited expression of required translational initiation factors such as eIF4E (Gingras *et al.*, 1999), or via specific interactions between proteins and either the untranslated region (UTR) or coding component of mRNA. The

expression of prokaryotic ribosomal protein operons is often controlled at the level of protein synthesis, where a product of the operon binds the transcript and prevents translation (Keener and Nomura, 1996). The CPEB protein in *Xenopus* prevents the progression of the cell cycle by binding a conserved sequence in the 3'-UTR of transcripts such as that of cyclin B1 (Cao *et al.*, 2002). CPEB recruits another protein, maskin, that interferes with the initiation of translation. Production of the cyclin B1 protein requires the destruction of CPEB (Mendez *et al.*, EMBO 2002).

Regulation is exerted through the destruction of transcripts as well as their creation. mRNA degradation in prokaryotes is often dependent on the presence or absence of secondary RNA structure such as stem-loops in the UTRs of transcripts. For instance, the 5′-UTR of the *E. coli* K12 *ompA* transcript more than triples its half-life in the cell, relative to modified transcripts that lack this feature (Kushner, 1996). In eukaryotes, cytoplasmic mRNAs are protected from exonuclease degradation by a guanosine cap at the 5′ end and a poly-adenylate tail at the 3′ end, both of which are bound by proteins (Hollams *et al.*, 2002). However, degradation can be incited through the binding of destabilizing proteins to '*cis*-elements' within the mRNA.

The control of transcription is a vital component in the regulation of gene expression in all three domains of life. While the details of transcriptional regulation and initiation differ between Archaea, Bacteria, and Eukarya, all three groups make extensive use of regulatory proteins to control the production of mRNA.

Transcription and Transcriptional Regulation

Common principles

The key element in transcription is RNA polymerase (RNAP), a multi-subunit complex that passes along one strand of double helical DNA (the 'template' strand) in the 3′ to 5′ direction and assembles a complementary strand of RNA. The assembly that includes RNA polymerase, the template DNA, and the nascent RNA strand is the *ternary complex* (Krakow and Fronk, 1969). The ternary complex assembles during the initiation

phase of transcription, where a multi-step procedure subject to several forms of regulation enables its formation (Reviewed in Beckett, 2001).

For initiation to occur, RNAP must localize to a short sequence of DNA called the *promoter*, which is typically found upstream of the coding sequence and adjacent to the start of transcription (Bautz and Bautz, 1970). RNAP binds directly to the promoter in Bacteria, but indirectly via a set of transcription factors (TFs) in Eukarya and Archaea (Struhl, 1999). This step in activation is the *closed complex*, because the double helical DNA has not yet been melted and the polymerase complex is not yet committed to transcription. For transcription to begin, RNAP must unwind the DNA near its binding site, thus creating a transcription 'bubble' where single-stranded DNA is exposed and can be read (Murakami *et al.,* 2002). If this step is successful, then the *open complex* containing RNAP and the unwound DNA is formed, and RNAP will begin to construct a chain of complementary ribonucleotides. The final step of initiation is *promoter clearance*, where the transcribing RNAP moves downstream from the promoter and initiation site and the promoter region anneals to return to its double-stranded state. Once RNAP has vacated the promoter region, another RNAP molecule may then bind and initiate transcription (Hsu, 2002; Dvir, 2002).

Constitutive transcription occurs if the initiation of transcription by RNAP is never blocked. However, *negative control* is often used to reduce the level of transcription initiation or to eliminate it entirely. Negative control is achieved through the actions of *repressor* or *silencer* proteins which bind conserved sequences in the DNA. These proteins block the initiation of transcription by interfering with the RNAP / promoter interaction, which can occur directly through physical occlusion of the promoter region, or indirectly by preventing the dissociation of other proteins that interfere with RNAP recruitment (Müller-Hill, 1998; Paszkowski and Whitham, 2001).

Even in the absence of repressors and silencers, transcription initiation may not readily occur at a promoter. A 'weak' promoter is a sequence of DNA whose structure is not strongly recognized and bound by RNAP or its accessory proteins, and is therefore inactive in the absence of other assisting factors (Wu *et al.*, 1992). *Positive control* allows initiation to occur at these promoters with the help of *activator proteins* that typically bind upstream of the promoter. Activator proteins can increase the level of transcription

from null or low levels, allowing the expression of genes that are normally inactive. As with negative regulators, activators can act directly by stabilizing the RNAP / promoter interaction through direct protein-protein contacts (Rhodius and Busby, 1998), or indirectly by inducing the removal of proteins that block transcription.

These positive and negative regulators are grouped together as *regulatory proteins*. A regulatory molecule will often be a homo- or hetero-oligomer, with DNA interactions dependent on the oligomerization state of the complex (Beckett, 2001). In addition to the oligomerization domain that permits complex formation, regulatory proteins also have a DNA-binding domain which allows them to bind their target regulatory sequence, and may also have a ligand-binding domain which allows them to interact with small signal molecules (see below).

The purpose of these regulatory proteins in transcription is to be responsive to changing conditions within the cell, to allow the rapid expression of specific genes only when they are needed. These proteins are often responsive to small *ligand* molecules in the cell such as trehalose (Arguelles, 2000), cyclic AMP (Botsford and Harman, 1992), and estrogen (McDonnell and Norris, 2002), which evoke changes in the state of regulatory proteins, either by inducing a conformational change, modifying the rate of oligomerization of regulatory protein components, or both. If the complex between regulatory protein and ligand is active and binds the target DNA sequence, then the regulatory protein is *inducible* by the ligand molecule. Conversely, if the ligand / protein complex is inactive, then the protein is *repressible*.

The activity and assembly of regulatory proteins can also be modified by post-translational methods. The σ^S protein of *E. coli* is required to initiate transcription of genes in response to cellular stress. Production of σ^S is regulated at the transcriptional and translational level, but its activity is also controlled through protein-protein interactions and through degradation. Degradation of σ^S requires the RssB protein, the activity of which appears to be mediated by the presence of small compounds such as acetyl phosphate (Hengge-Aronis, 2002). A set of proteins called *anti-sigma factors* can also prevent sigma factor activity. An example of this effect is the RseA/RseB protein complex in *E. coli*, which can bind and inactivate σ^E, a sigma factor that initiates transcription of genes in response to accumulation of unfolded proteins in the cell. σ^E is

only released if denaturing cellular conditions (such as extreme heat) cause the dissociation of RseB from the complex (Tam *et al.*, 2002).

The protein-DNA and protein-protein interactions that occur during gene expression are governed by rates of association, dissociation and reaction. However, since the number of participants in a given reaction within a cell is typically quite small, there appears to be a random element in the regulation of gene expression. McAdams and Arkin (1997) performed simulations that showed the potential for substantial differences in protein concentrations between different cells within a clonal population. These differences are due to random differences in the association rate of RNA polymerase with specific promoters, competitive binding of ribosomes and RNAses to transcripts, and association rates of monomers to yield active regulatory protein complexes.

While the principles outlined above are common to all three domains of life, the strategies used to implement these principles are very different. In the broadest sense, initiation of transcription can be separated into two components that highlight the differences between domains. The mechanism of transcription initiation is similar and likely homologous between Archaea and Eukarya, with both systems dependent on accessory factors to allow RNAP / promoter interactions. Bacteria possess a much simpler initiation mechanism, with direct contacts between RNAP and DNA (Murakami *et al.,* 2002). However, the mode of transcriptional regulation is similar between Archaea and Bacteria, with repressors and activators binding near the promoter to modify the rate of transcription. In contrast, Eukarya rely primarily on transcriptional activator proteins, many of which bind to enhancer regions far from the promoter: this bias toward positive regulation exists because of the frequent sequestering of eukaryotic DNA by histone proteins that block the promoter.

Bacteria

While components of the transcription and regulation machinery have been studied in many bacterial genomes, the majority of information about bacterial regulation has been derived from two species: the enteric bacterium *Escherichia coli* and, to a lesser degree, the Gram-positive *Bacillus subtilis*. These organisms illustrate transcriptional

features common to many studied bacteria, including RNA polymerase sigma factors, promoters of the form (conserved domain – spacer – conserved domain), and operon organization, which are described below.

In bacteria, the RNA polymerase *core enzyme* is responsible for the elongation of a transcribed RNA molecule. The core enzyme has five constituent subunits: a homodimer containing two identical α proteins, two positively charged proteins, β and β', and the small subunit ω. The overall structure resembles a 'crab claw' (Darst, 2001) with β and β' forming the pincers of the claw. The α homodimer connects the β and β' subunits, and ω is peripheral at the base of the claw. The cleft formed by the β and β' subunits is where catalysis occurs, while the α proteins provide structural and regulatory functions, and ω assists in assembly, specifically by assisting in the attachment of β' to the rest of the core (Minakhin *et* al., 2001).

The core enzyme cannot initiate transcription on its own. Instead, the σ subunit must associate with the core enzyme prior to transcription, yielding the *holoenzyme*. The σ subunit alone cannot bind to DNA, but binding to the β and β' subunits of the core enzyme yields a conformational change that exposes the DNA-binding domains of σ (Dombroski *et al.,* 1993; Nomura *et al.,* 1999; Katayama *et al.,* 2000). These DNA-binding domains then localize to the promoter and bind it directly, typically at two short (~6 nucleotide) sequences centered around positions –10 and –35 relative to the start of transcription, that are separated by a longer, nonconserved spacer region (Murakami *et al.,* 2002). The σ factor then uses its anchor points to twist apart the DNA, yielding the open promoter complex (Naryshkin *et al.*, 2000). Once σ has performed this task, it dissociates from the core enzyme, allowing the core enzyme to begin transcribing the template sequence (Daube and von Hippel, 1999). Sigma factors have several conserved domains, including regions that are responsible for –10 recognition ($\sigma_{2.3}$ and $\sigma_{2.4}$ in the crystallized *Thermus aquaticus* σ^A protein), -35 recognition (σ_4), and a region that appears to recognize "extended –10" promoters which lack a conserved –35 region but have more conserved nucleotides near the –10 box (Young *et al.*, 2002).

Transcription is often aborted after formation of a few (< 10) phosphodiester bonds between ribonucleotides, leading to release of the short RNA molecule and another attempt at initiation. The rate of abortive release appears to be somewhat dependent on

the sequence of the promoter, with release more frequent at stronger promoters due to the difficulty in breaking the RNAP / promoter interactions (Hsu, 2002). These release events always occur prior to the dissociation or repositioning of σ factor, and the release of σ, which usually occurs when the nascent transcript is 9 or 10 nucleotides in length, signals the end of initiation and the start of elongation (Susa *et al.*, 2002). Structural studies support this view by showing that a subdomain of σ blocks the 'tunnel' in RNAP where ribonucleotide assembly is performed (Hsu, 2002).

Co-expression of bacterial and archaeal genes is often achieved through *operon* organization. In an operon, several protein-coding sequences are arranged in the same direction, with very small (< 50 nucleotide) spacers between open reading frames. Operonic genes are transcribed together, and are under the control of a single promoter upstream of the first operonic gene (Beckwith, 1996). Each protein-coding sequence within an operon has its own Shine-Dalgarno ribosome binding site (Shine and Dalgarno, 1974) and is translated independently from the rest of the open reading frames. This transcriptional strategy is frequently used to co-regulate expression of proteins in biosynthetic pathways for amino acid synthesis (Yanofsky, 2000), nutrient catabolism (Reznikoff, 1992), and ribosomal proteins (Keener and Nomura, 1996).

Negative control in *E. coli* and other bacteria is achieved through repressor protein complexes. These repressors bind a recognition sequence that is typically just downstream of the promoter, and prevent either RNAP binding or initiation (Müller-Hill, 1998). The LacI protein can block transcription at either of these steps: for instance, by preventing the RNAP / promoter interaction in the *lacZYA* operon, or by forcing abortive release in the *lacUV5* operon (Hsu, 2002; Straney and Crothers, 1987). A tetrameric LacI protein actually binds two distal operators, causing the DNA to loop around and block interactions with CAP (see below), thus preventing transcriptional activation (Müller-Hill, 1998). Another well-studied repressor in *E. coli* is TrpR, which binds upstream of several operons that encode tryptophan biosynthetic proteins: *trpEDCBA*, *trpR*, *mtr,* and *aroH*. Up to three TrpR homodimers can bind the *trpEDCBA* and *aroH* operators at a single time, while the other operators can interact with two TrpR dimers (Jeeves *et al.*, 1999). TrpR is active in the presence of tryptophan, but dissociates when tryptophan is absent, allowing transcription to proceed. Another level of sensitivity to tryptophan levels

is achieved through *attenuation*, where the progress of protein synthesis through a pair of Trp codons in the *trpL* gene can cause premature termination of transcription (Yanofsky, 2000).

Repressor proteins can prevent transcription in an indirect manner *via* DNA looping. In the absence of arabinose, the AraC protein from *E. coli* binds a pair of recognition sites upstream of the *araBAD* operon that are separated by 210 nt. These interactions cause the upstream DNA to form a loop, which prevents the binding of RNA polymerase to the *araBAD* promoters located within the looping region (Schleif, 2003).

The bacterial nucleoid consists of the bacterial chromosome complexed with a number of proteins and nascent RNA chains (Pettijohn, 1996). In *E. coli*, the chromosome assumes a 'rosette' structure, with a compact center and ~50 projecting loops. Several proteins stabilize this structure, including HU, IHF, H-NS and FIS. These nucleoid proteins perform several functions, including DNA compaction, partitioning the chromosome into independent supercoiling domains, and anchoring the bacterial DNA to the cell membrane (Ali Azam *et al.*, 1999*a*; Pettijohn, 1996). While they have different functions and DNA-binding properties, mutants for one type of nucleoid protein are often viable, with other proteins filling in some (but not all) of the missing protein's functions. For instance, the detrimental effect of a mutant IHF subunit can be compensated with homodimers of the other subunit, as well as HU (Bykowski and Sirko, 1998). However, cells with mutant nucleoid proteins are very sensitive to some types of stress such as ultraviolet irradiation (Li and Waters, 1998). There are a total of 15-20 nucleoid proteins, with a few such as IHF, HU, H-NS and Fis designated as 'major', and several others, including Lrp and StpA, termed 'minor' (Ali Azam *et al.*, 1999*b*).

The nucleoid proteins also play important roles in the regulation of transcription. Integration host factor (IHF) binds DNA and induces bends of up to 180°, yielding a U-turn structure that can bring proteins with binding sites ~50 nt apart into close proximity (Travers, 1997). The consensus sequence for IHF binding is known, but the key to IHF binding and distortion is the conformational mobility of the base pair steps to be unwound (Steffen *et al.*, 2002). The reliance on structural properties of DNA rather than direct interactions with specific nucleotide functional groups is termed *indirect* recognition (Travers, 1997).

The HN-S protein and its presumed paralog StpA are also implicated in transcriptional regulation. These proteins can form homo- or heteromeric complexes, which can constrain DNA supercoils and repress transcription from some promoters. While the *bgl* operon appears to be sensitive to the levels of both H-NS and StpA, suggesting a role for a heteromeric complex, the *proU* operon appears to be regulated by H-NS alone (Dorman *et al.*, 1999).

Another important nucleoid protein in *E. coli* is HU. Like IHF, active HU is a heterodimer, but HU does not recognize DNA specifically. Instead, HU appears to preferentially bind bent DNA, stabilizing or enhancing this conformation to permit further activity (Wojtuszewski *et al.*, 2001). HU has been implicated in the activation of transcription at the *lac* promoter by assisting with the binding of cAMP/CAP and LacI (Flashner and Gralla, 1988). Other roles for HU include the stabilization of supercoiling domains through binding of four-way junctions (Pettijohn, 1996).

The last of the major nucleoid proteins is Fis. The role of this protein is primarily in the control of supercoiling stability, which is effected primarily through antagonism of DNA gyrase. Fis interferes with gyrase production by inactivating the transcription of the *gyrA* and *gyrB* genes, and through interactions with DNA that decrease the effectiveness of gyrase proteins (Schneider *et al.*, 1999). The expression of *fis* is itself tied to the supercoiling state of the genome: its promoter is most active if negatively supercoiled, and expression of *fis* is decreased dramatically if the degree of supercoiling increases or decreases. Thus, it appears that Fis may be responsible for maintaining the optimal supercoiling density required for gene expression during exponential growth (Schneider *et al.*, 2000).

Positive control is most frequently exerted through activator proteins which bind between 70 and 120 base-pairs upstream of the transcription start site (Rhodius and Busby, 1998). The activators can contact either the carboxy-terminal domain (CTD) or N-terminal domain (NTD) of the α subunit, or a domain of the σ factor, or both. This can stabilize the interactions with weak promoters, increase the rate of open complex formation, or assist in promoter escape (Monsalve *et al.*, 1996). The cyclic AMP activated protein (CAP) is required for transcriptional activation of many genes, including those responsible for non-glucose sugar degradation and pyrimidine

metabolism (Botsford and Harman, 1992). CAP forms a homodimer complex, and only binds DNA when in contact with cyclic AMP. At different promoters, the CAP binding site can be distal, or proximal, or both. Distal binding of CAP assists in closed complex formation by increasing the affinity of RNAP for the promoter *via* an interaction between CAP and the α-CTD of RNAP. Proximal binding increases the rate of open complex formation through interactions with α-NTD, though the mechanism of this effect is not yet known (Busby and Ebright, 1999).

While activator proteins that bind several hundred base pairs upstream of the promoter are more characteristic of eukaryotic regulation (see below), some bacterial activator proteins can act at a distance to permit the initiation of transcription. Open complex formation by RNAP complexed with σ^{54} requires the assistance of an activator protein such as NtrC. NtrC, which is activated through phosphorylation, has been shown to activate transcription at the *E. coli* K12 *glnA* gene even when its binding site is placed 3000 base pairs upstream of the promoter. The contact between NtrC and RNAP is enabled through DNA looping, and can be influenced by the structure of the intervening DNA (Schulz *et al.*, 2000; Xu and Hoover, 2001). DNA looping can be induced through the binding of the integration host factor (IHF) protein, which often binds between the activator site and the promoter, and induces a 160° bend in the bound DNA (Claverie-Martin and Magasanik, 1991).

Some regulatory proteins can act either as activators or repressors. The TyrR protein can repress the transcriptional activation of genes required in aromatic amino acid synthesis, but is also required for activation of transcription at several genes including *mtr*. The *aroP* gene has three promoters, which together illustrate the activation and repression functions of TyrR. Transcription of *aroP* can be initiated from promoters P1 and P2, while P3 is oriented in the other direction, away from the coding sequence. When tyrosine and other cofactors are absent, TyrR binds a 'strong box' that reduces *aroP* expression by interfering with steps after open complex formation (Yang *et al.*, 1999). The same bound TyrR complex activates transcription from the divergent P3 promoter, where the stabilized bound RNAP sterically prevents activation from P1 (Wang *et al.*, 1997).

While the σ^{70} protein in *E. coli* is the constitutive or 'vegetative' sigma factor, other sigma factors are required to express sets of genes in response to changing cellular conditions. Six alternative σ factors are known in *E. coli*: σ^{S} for stationary phase genes, σ^{32} for heat-shock genes, σ^{F} for flagellar growth, σ^{E} for extreme heat shock and ethanol stress, σ^{fecI} for ferric citrate transport, and σ^{N} for nitrogen assimilation and metabolism (Reitzer and Schneider, 2001; Maeda *et al.*, 2000; Wosten, 1998). *Bacillus subtilis* has 16 known alternative sigma factors (Kroos and Yu, 2000) in addition to its vegetative (σ^{A}) protein, six of which are required for a regulatory cascade that occurs during sporulation. The increased expression and activity of *E. coli* σ^{32} during the heat shock response is due not to modulated transcription levels, but to increased translation resulting from heat-induced structural modifications in the mRNA (Morita *et al.,* 2000). After several minutes, σ^{32} is bound by the DnaJ and DnaK proteins and probably targeted for protease cleavage (Ramos *et al.*, 2001). The expression of other sigma factors can occur in response to small ligand molecules, such as trehalose and ppGpp (Cases and de Lorenzo, 1998).

Coordinated response to cellular stress can also be achieved with the stringent response. In *E. coli*, the presence of unconjugated transfer RNA molecules can indicate nutrient shortage. Under such conditions, the RelA protein catalyzes the reaction of guanosine diphosphate with pyrophosphate to yield the molecule ppGpp (Cashel *et al.*, 1996). This molecule appears to interact with the β subunit of RNA polymerase, inducing a conformational change that modifies the binding preference of RNAP. This modified preference leads to decreased expression of some genes, particularly those encoding ribosomal and transfer RNAs, and increases the transcription rate of other genes such as *rpoS*, which encodes a general stress-response sigma factor (Chatterji and Ojha, 2001).

Eukarya

Several model eukaryotic organisms have been extensively studied, including the budding yeast *S. cerevisiae* (Goffeau *et al.*, 1996; Foury *et al.*, 1998), the fission yeast *Schizosaccharomyces pombe* (Wood *et al.*, 2002), *Drosophila melanogaster* (Adams *et al.,* 2000) and *Homo sapiens* (Lander *et al.,* 2001). While the regulatory proteins used in

the same type of response may differ between eukaryotic organisms, the basic mode of transcriptional regulation is conserved throughout the domain.

Unlike bacteria, where a single RNA polymerase core enzyme carries out all transcription duties, eukaryotes have three different multi-subunit polymerases, identified with Roman numerals: RNA polymerase I (pol I) transcribes most ribosomal RNA genes, RNA polymerase II (pol II) transcribes protein-coding genes, and RNA polymerase III (pol III) transcribes 5S rRNA, all nuclear tRNAs, and other small RNA genes (Paule and White, 2000). All three polymerases are far more complex than bacterial RNAP, consisting of at least twelve subunits each (Sentenac *et al.,* 1992), five of which (Rpb1, Rpb2, Rpb6, and Rbp3-Rpb11 in yeast pol II) are homologous to the bacterial RNAP core subunits (respectively, β', β, ω, and α-α). (Minakhin *et al.*, 2001). Other subunits found in pol I, II, and III are involved in start site selection (Rpb9), interactions with activator proteins (Rpb5), or response to certain nutrient conditions (Rpb4 and Rpb7) (Ishihama *et al.*, 1998). While pol I and pol III are responsible for the majority of transcripts produced (Paule and White, 2000), most eukaryotic genes are transcribed by pol II.

Eukaryotic RNA polymerases do not bind directly to the DNA. Instead, a set of proteins called the *general transcription factors* (GTFs) must first assemble on the promoter (Burley and Kamada, 2002). Six multiprotein complexes (TFIID, TFIIB, TFIIA, TFIIF, TFIIE, TFIIH) are involved in pol II recruitment to a promoter, and formation of the promoter initiation complex. A component of the TFIID complex, TATA-binding protein (TBP) binds the promoter DNA and allows the recruitment of the other GTFs to the promoter site (Lagrange *et al.*, 1998). While not all subunits have yet been associated with a function, components of some GTFs are known to assist in DNA binding, activator interactions, polymerase recruitment and stabilization, promoter opening, and start site selection (Lee and Young, 2000) TFIIH is essential to promoter clearance, and is involved in transcript proofreading and repair (Eisen and Lucchesi, 1998). Eukaryotic RNA polymerases are susceptible to the same type of abortive release as their bacterial counterparts; TFIIF is involved in suppressing this event (Dvir *et al.*, 2001).

The binding of the RNAP complex induces extensive change in DNA conformation. Crystal studies of TBP/promoter complexes show that TBP bends the TATA box at an angle of ~90° (Kim *et al.*, 1993*b*). The binding of sequence elements such as TFIIH upstream and downstream of the transcription start site results in the tight wrapping of 50-90 bp of DNA around the polymerase complex prior to the initiation of transcription (Eisen and Lucchesi, 1998).

The key difference between bacterial and eukaryotic regulation is the presence of nucleosomes in eukaryotic DNA. Histone octamers bind and sequester coding and upstream regions, preventing transcription initiation and elongation (Bauer *et al.*, 1994). Therefore, most eukaryotic genes are inactive by default, and activator proteins are required to allow chromatin displacement and initiation complex assembly (Struhl, 1999). Two enzymes are required to disrupt the histones that block promoter access. The Swi/Snf complex, which has a single representative and a limited role in yeast, but several types and a more extensive role in humans, can either inactivate histones through remodeling (Sudarsanam and Winston, 2000), or through displacement, in either *cis* or *trans* (Peterson and Workman, 2000; Schnitzler *et al.*, 1998). The Swi/Snf complex in yeast and a human counterpart SWI/SNF have been implicated in both transcriptional activation and repression, because in remodeling chromatin, they facilitate DNA binding of repressor proteins as well as activating transcription factors (Sudarsanam and Winston, 2000). Histone acetyltransferases (HATs) can disrupt nucleosome structure through acetylation of the basic histone N-terminal tails (Horn and Peterson, 2002; Strahl and Allis, 2000; Kuo and Allis, 1998). Since these positively-charged tails are implicated in DNA binding, interactions with silencer proteins, and higher-order histone-histone interactions, the addition of an acetyl group can disrupt all of the above functions (Luger and Richmond, 1998*b*). Several HAT proteins have been implicated in yeast transcriptional activation:, including Gcn5p, PCAF, and the TFIID component $TAF_{II}250$ (Kuo and Allis, 1998).

Chromatin remodeling can be induced through activator proteins, and repressed by *silencer* proteins. Transcriptional silencing is observed in yeast at the telomeres, centromeres, pol I-transcribed ribosomal DNA genes, and at the mating-type (HM) loci. Silencing at the HM loci is mediated by the Sir complex of proteins, which interacts with

silencer elements in nearby DNA via three intermediate proteins: ORC, Rap1p, and Abf1p (Huang, 2002). These proteins bind specific DNA sequences, and then recruit the Sir proteins. The mechanism of transcriptional silencing by the Sir complex is still unclear: interestingly, recent studies suggest that silencing does not necessarily preclude the formation of the pre-initiation complex (Sekinger and Gross, 2001).

Silencing is also achieved in some eukaryotes through methylation of upstream CpG sequences (Robertson, 2002; Yoder *et al.*, 1997). While upstream CpG dinucleotides are usually hypomethylated with respect to other regions of the genome, defective DNA methyltransferases can lead to hypermethylated upstream regions, with the resulting disruption of gene expression linked to cancer (Baylin *et al.,* 2001; Jones and Laird, 1999; Robertson, 2001). Methylated CpG regions are typically bound by proteins that deacetylate nearby histones, leading to a lesser likelihood of histone remodeling (Robertson, 2002).

In yeast, the Swi/Snf remodeling complex is recruited by activator proteins such as Gal4, Gcn4, Hap4 and VP16 (Neely *et al.,* 1999; Natarajan *et al.*, 1999; Yudkovsky, 1999). Such activator proteins typically bind upstream activating sequences (UASs) in yeast, or enhancer elements in metazoans that need not be proximal to the promoter (Hampsey, 1998). UASs and enhancers tend to organize into regulatory elements of between 50 and 1500 base pairs in length, with a single regulatory element typically conferring activation in a specific cell type, developmental stage, or cellular condition (Blackwood and Kadonga, 1998). DNA looping, due in part to the compacted chromatin structure, allows protein-bound enhancers to interact with the promoter region (Ogata *et al*, 2003). Specific combinations of activator proteins and cofactors are recruited to different regulatory elements; an active combination that yields transcriptional activation is termed the *enhanceosome* (Merika and Thanos, 2001). These enhanceosomes can assist in the initiation of transcription through recruitment of chromatin-altering proteins (Sudarsanam and Winston, 2000), and phosphorylation of transcription factors such as TFIIE, TFIIF as well as the RNA polymerase (Ishihama *et al.*, 1998).

An example of a eukaryotic activator is the product of the human *c-myc* proto-oncogene. The Myc protein product of this gene is a critical regulatory component of the cell division process and of apoptosis, and mutations that affect the activity or expression

of Myc can lead to unregulated cell proliferation and cancer (Amati *et al.*, 2001). Myc interacts with another protein, Max, to form a heterodimer that binds a short enhancer element. Myc has been implicated in the recruitment of several HATs (McMahon *et al.*, 2000) and Swi/Snf complexes (Cheng *et al.*, 1999), supporting the role of Myc as an inducer of chromatin modification. The Myc protein has also been implicated in down-regulation of some genes, but this appears to result indirectly from increased expression of repressors that are directly targeted by Myc (Claassen and Hann, 1999).

Upstream repressor sequences (URSs) are the repressive version of UASs. Repressor proteins bound to URSs can block the binding of activators, interfere with the function of bound activators, and stabilize histone-DNA interactions through deacetylation (de Ruijter *et al.*, 2003). Mad is a human protein that heterodimerizes with the Max protein and binds the same recognition sequence as the Myc-Max dimer, thus creating a competition for binding (Luscher, 2001). Mad is responsible for the recruitment of histone deacetylases which lead to repression of transcription (Amati *et al.*,2001), thus providing a negative 'switch' for chromatin remodeling that complements the activity of Myc.

The *promoter-proximal region* is located between 200 and 50 base pairs upstream of the transcription start site, adjacent to the core promoter region (Roy and Lee, 1995). A few well-characterized proteins are known to bind specifically in this region, many interacting with a sequence containing a core CCAAT sequence. CCAAT-binding transcription factors include CCAAT-enhancer binding protein (c/EBP), CCAAT transcription factor (CTF), and nuclear factor Y (NF-Y) (Mantovani, 1998). NF-Y is involved in the regulation of ~25% of all eukaryotic promoters (Zemzoumi *et al.*, 1999), and is found preferentially in TATA-less promoters. The NF-Y binding site tends to be close to the transcription initiator in TATA-less promoters, but further upstream in promoters that contain a TATA box (Mantovani, 1998). Sp1 is another common transcription factor, which appears to be constitutively expressed but regulated through posttranslational modification and preferentially active in certain cell types especially in the regulation of growth (Black *et al.*, 2001). The Sp1 family of proteins bind recognition sequences such as the GC box (consensus GGGGCGGGC) and GT box (consensus GGGTGTGGC) (Hirano *et al.*, 1998).

Repression can also be achieved without DNA binding. Certain proteins can interfere with the function of critical transcription factors and preventing DNA binding or oligomerization. The yeast protein Mot1 specifically targets the TBP/DNA complex and forces the dissociation of TBP. The Hsp90 chaperone binds the transcriptional activator Hsf1, preventing the formation of active trimers (Lee and Young, 2000).

Archaea

The Archaea represent a heterogeneous group of organisms, with significant differences in regulatory features. Homologues of eukaryal histone proteins are found in Euryarchaeota but not Crenarchaeota (Reeve *et al.*, 1997), leading to different regulatory needs in these two divisions.

Despite the diversity in Archaea, a single >10 subunit RNA polymerase has been identified (Huet *et al.*, 1983), in euryarchaeotes such as *Methanobacterium thermoautotrophicum* and *Methanococcus jannaschii*, and crenarchaeotes including *Sulfolobus solfataricus* (Bell and Jackson, 2001). The three-dimensional structure of archaeal RNAP is very similar to those of Bacteria and Eukarya, and the subunit organization is very similar to its eukaryotic counterpart (Cramer *et al.*, 2001). In particular, two components of archaeal RNAP, A′ and A′′, are homologous to the bacterial β' subunit and the corresponding eukaryal Rbp1, and the counterparts of the bacterial β subunit are the B′ and B′′ components (Langer *et al.*, 1995). The archaeal F and P subunits are homologous to the eukaryotic Rpb4 and Rpb12 (Bell and Jackson, 2000; Werner *et al.*, 2000), and the archaeal RpoK protein is homologous to the bacterial ω subunit and the eukaryotic Rpb6 gene product (Magill *et al.*, 2001). Sequence and structural analysis have shown that the core architecture and function of RNA polymerase is consistent between all three domains of life (Cramer *et al.*, 2001), with the strongest similarity visible between archaeal RNAP and eukaryotic Pol II (Magill *et al.*, 2001).

RNAP recruitment at the promoter is similar in Archaea and Eukarya, though the archaeal mechanism requires fewer intermediates. The archaeal homologue to eukaryotic TBP is also called TBP, which also binds a conserved TA-rich promoter sequence (Qureshi and Jackson, 1998), but archaeal TBP does not interact with other proteins to

yield a TFIID-like complex (Hausner *et al.*, 1996; Bell *et al.,* 1998). Instead, archaeal TBP is directly responsible for the recruitment of transcription factor B (TFB), which then interacts with the RpoK subunit of RNAP and ensures the correct orientation of the enzyme through interaction with a 'B response element' or BRE (Magill *et al.*, 2001). The only other general transcription factor identified thus far in Archaea is TFE, which is thought to play a stimulatory but non-essential role in activating transcription (Bell and Jackson, 2001; Qureshi and Jackson, 1998). It appears that archaeal TBP and TFB may also play a role similar to bacterial σ: multiple variants of these two proteins have been found in sequenced archaeal genomes, and the expression of one TFB variant in *Haloferax volcanii* is associated with the heat-shock response (Bell and Jackson, 2001). Alternative TFBs appear to confer different binding affinities for different BRE sequences (Baliga and Dassarma, 2000; Thompson *et al.,* 1999; Thompson and Daniels, 1998).

As with the other two domains of life, Archaea use proteins to bind and sequester their genomic DNA. However, the Archaea are much more heterogeneous in the identity and function of these proteins, which are specific to sub-groupings within the Archaea. Taxa within one of the two major taxonomic divisions, the Euryarchaeotes, contain one or more histone-like proteins, which are structurally very similar to the core folds of the eukaryotic H3 and H4 nucleosome components (Luger and Richmond, 1998*a*). Histone proteins in Archaea form either a homo- or heterotetramer, and there is some evidence that different combinations of histone proteins can have greater or lesser repressive effects on transcription (White and Bell, 2002). Little is yet known about histone remodeling or displacement in archaeal transcriptional activation.

A great diversity of other DNA compaction strategies exist within the Archaea. One of the best-studied of these is the Alba protein, which appears to have a repressive effect on transcription. In *Sulfolobus*, Alba is often found in complex with a homologue of the eukaryal Sir histone deacetylase, which can yield extensive transcriptional repression (Bell *et al.,* 2002). Alba appears to bind the DNA without yielding much compaction, but genomes that contain Alba also contain one of several other types of compaction proteins (White and Bell, 2002), suggesting a cooperative role in compacting DNA and repressing transcription.

Despite the presence of histone-like proteins, archaeal regulation of transcription initiation is similar to Bacteria in nature (Kyrpides and Ouzounis, 1999). Negative control is achieved through the action of repressor proteins. A well-studied example is the family of autoregulatory leucine-responsive regulatory proteins (Lrps) found in Archaea. These proteins bind at or near the promoter, and interfere either with recruitment of the general transcription factors (Bell and Jackson, 2000) or with polymerase recruitment (Dahlke and Thomm, 2002).

A small number of activator proteins have been identified in Archaea. The GvpE protein in haloarchaea activates the transcription of an operon of genes responsible for gas vesicle synthesis (Kruger *et al.*, 1998). The Bat gene encodes an activator protein required for the synthesis of bacteriorhodopsin in the extreme halophile *Halobacterium* sp. NRC-1, and a consensus for its UAS, located within 7 base-pairs of the promoter, was identified (Baliga *et al.*, 2001). The mode of action of these activator proteins has not yet been elucidated.

DNA Sequence and Structure

Properties of DNA

In a sequence, the four nucleotides of DNA are usually represented with their IUPAC designations (Table 1-1). Since transcription factors and bacterial RNAP can generally bind to more than one sequence of DNA, the IUPAC nomenclature is useful in representing the allowed degeneracy of protein-binding sites. In particular, twofold degenerate sites can represent a conserved nucleotide size (puRine vs. pYrimidine), number of hydrogen bonds (Strong = 3 bonds, Weak = 2 bonds) or functional group (Keto or aMino).

While structural variations in double helical DNA are less dramatic than those observed in proteins, the different properties of the four nucleotide bases do confer variability in the shape and flexibility of the DNA molecule. A number of basic structural parameters are used to describe the translations and rotations between adjacent base pairs and between paired nucleotides. The x-axis of a base pair points along the short axis, the

Table 1-1: IUPAC codes for nucleotide bases.

IUPAC Letter	Description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
R	Purine (A or G)
Y	Pyrimidine (C or T)
S	Strong (C or G)
W	Weak (A or T)
K	Keto (G or T)
M	Amino (A or C)
B	Not A (C, G, or T)
D	Not C (A, G, or T)
H	Not G (A, C, or T)
V	Not T (A, C, or G)
N	Any (A, C, G, or T)

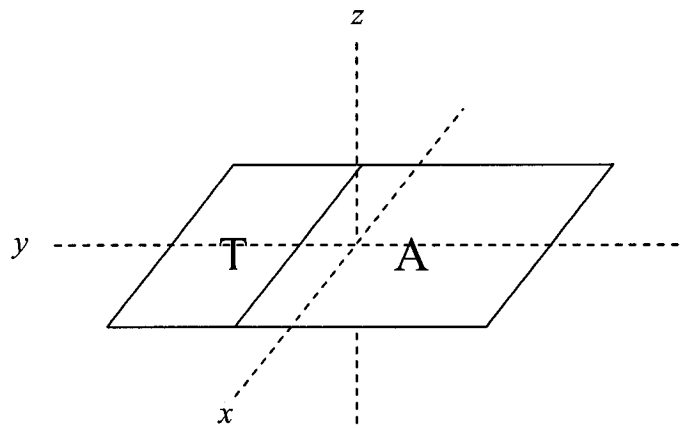
y-axis is parallel to the long axis of a base-pair, and the z-axis is perpendicular to the plane of the base pair (Figure 1-1). The Cambridge Accord (Dickerson, 1989) formalized the rotations and translations along these axes, which are summarized in Table 1-2. Other properties of the DNA such as curvature and major and minor groove width can be derived if the above structural parameters are known for a given molecule of DNA (Gorin *et al.*, 1995). In addition to these static structural features, the conformational mobility of DNA can also be described with several quantities, including persistence length (Vologodskaia and Vologodskii, 2002), flexibility (Olson *et al.*, 1998), and deformation energy (Steffen *et al.*, 2002).

While the Sanger chain termination reaction (Sanger *et al.*, 1977) has been automated and used in many high-throughput sequencing projects, the determination of DNA structure at the atomic level is very difficult. Early fiber diffraction studies allowed the identification of the different alternative stable helix types (A-DNA, B-DNA, and Z-DNA: Dickerson *et al.*, 1982), but conclusions about structure were based on the mean parameter values of the entire helix, and the contributions of individual base pairs to the overall structure could not be assessed. Similarly, large-scale DNA curvature was assessed in the 1980s using comparisons of electrophoretic mobility. The presence of long adenine tracts was associated with greater intrinsic curvature, but these assays were unable to distinguish between three alternative hypotheses for A-tract-dependent binding: bending within A-tracts, bending at A/non-A junctions, or non-A-tract bending spaced by rigid A-tracts (Dickerson *et al.*, 1994).

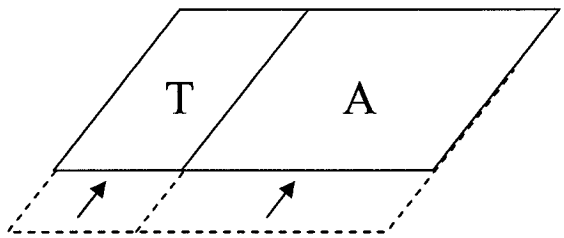
Structural determination of DNA at the atomic level was first achieved with crystallization experiments. In the 1970s, crystal structures of oligonucleotides were obtained, starting with a dinucleotide of double-stranded RNA in 1973 (Rosenberg *et al.*, 1973), followed by oligonucleotides that showed the Z conformation (Drew *et al.*, 1980). A major landmark in the field was the resolution of the 'Drew-Dickerson dodecamer', a self-complementary B-DNA double helix with the sequence d(CGCGAATTCGCG) (Wing *et al.*, 1980). Since then, a large number of B-DNA dodecamers and decamers have been solved and deposited at the Nucleic Acid Database (Berman *et al.*, 2003), including naked B-DNA, DNA with mismatches, DNA-drug complexes, and protein-DNA complexes. The structural parameters of these molecules have been summarized in

Figure 1-1: Axis definitions (*a*) and illustrations of co-ordinated (*b*) and opposed (*c*) base pair transformations.

a) Axes



b) Co-ordinated (*x*-displacement)



c) Opposed (Shear)

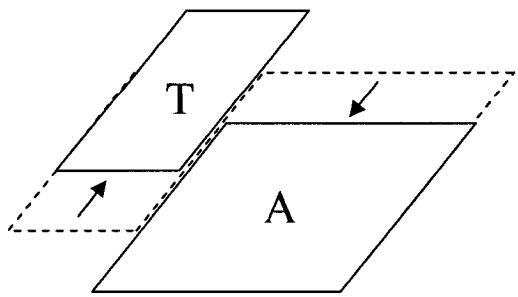


Table 1-2: Rotations (a) and translations (b) of nucleotide base pairs, with associated symbols. N/A, not applicable.

a) Rotations

Axis	Base pair co-ordinated	Base pair opposed	Adjacent base pairs
X	Inclination (η)	Buckle (κ)	Tilt (τ)
Y	Tip (θ)	Propellor Twist (ω)	Roll (ρ)
Z	N/A	Opening (σ)	Twist (Ω)

b) Translations

Axis	Base pair co-ordinated	Base pair opposed	Adjacent base pairs
X	x-displacement (dx)	Shear (Sx)	Shift (Dx)
Y	y-displacement (dy)	Stretch (Sy)	Slide (Dy)
Z	N/A	Stagger (Sz)	Rise (Dz)

several reports (Gorin *et al.,* 1995; el Hassan and Calladine, 1996), and interpreted in light of theoretical models (Hunter 1993). Every possible dinucleotide is present many times in the data set, and most of the structural summaries deal with these values. While trends have been identified in the empirical data and supported by theory, a dinucleotide model is clearly too simplistic to represent the structure of DNA accurately (Yanagi *et al.,* 1991; Dickerson *et al.,* 1994).

While crystal studies have highlighted important sequence-specific features of DNA, the biases in this data set limit the relevance to structural understanding of DNA *in vivo*. An important concern is the bias in the oligonucleotides whose structures have been determined thus far: many of them are very similar to the Drew-Dickerson dodecamer, and the majority have CG dinucleotides at both ends. Since genomic DNA is often circular and is longer than 12 nucleotides, its structure is constrained in ways that the crystallized DNA molecules are not (Neidle, 1996). Finally, the influence of the crystal environment itself is controversial: the crystallographic conditions of most solved B-DNA molecules are very similar, with many solution components and concentrations nearly identical. Lattice packing has been shown to influence DNA conformation preferentially at some steps (DiGabriele *et al.*, 1989), which is a caveat for interpretation of structures, but can be exploited as a source of information about sequence-dependent flexibility as well (Dickerson *et al.,* 1994). For instance, adenine-containing tracts appear to be rigid and minimally affected by crystal packing forces, suggesting that DNA bending occurs outside of these regions. This observation is supported by theoretical modeling of A-tract parameters such as propellor twist (el Hassan and Calladine, 1996).

An important aspect of DNA structure and flexibility supported by empirical and theoretical data is the vital role of nucleotide exocyclic groups such as the amino groups of adenine and cytosine, and the methyl group of thymine. Gorin *et al.* (1995) showed that the mean values of Twist, Slide, Rise, and Roll for each dinucleotide were strongly correlated with their 'clash function' that reflected the size of the exocyclic groups of the adjacent dinucleotides. El Hassan and Calladine (1996) showed that propellor twist was an important determinant of dinucleotide flexibility, because the opposed y-axis rotation of bases increases the likelihood of steric clashes with neighbouring base pairs, thus reducing the available space for conformational variation. The rigidity of A-tracts is due

to the high tendency of A-T pairs to undergo propellor twisting, and the high chance of steric clash between adjacent adenines. The conformational variability of pyrimidine/purine steps, identified originally in crystal data has more recently been verified with nuclear magnetic resonance studies (Kojima *et al.*, 2001; Derreumaux and Fermandjian, 2000)

Supercoiling is also an important property of DNA *in vivo*. Modulation in DNA supercoiling has been implicated in compacting of DNA (Stuger *et al.,* 2002), changing the propensity for reactions such as transcription initiation (Opel *et al.*, 2001), and increasing or decreasing site-specific protein-DNA recognition (Baliga *et al.,* 2001). Topoisomerases are responsible for modifying the degree of supercoiling, and *E. coli* has been shown to organize its DNA into multiple independent 'supercoiling domains' (Sinden and Pettijohn, 1981). The DNA of hyperthermophilic Bacteria and Archaea tends to be positively supercoiled through the actions of reverse gyrase, conferring resistance to temperature-induced denaturation (Lopez-Garcia, 1999).

Regulatory proteins

The successful association of proteins with DNA depends on the establishment of contacts between structural motifs in the protein and features of the DNA. The structure of many protein-DNA complexes have now been determined (Huffman and Brennan, 2002; Pabo and Nekludova, 2000), and several distinct classes of DNA-binding motif have been identified. DNA-binding proteins can interact with the target DNA via binding and hydrogen bonding in the major groove, in the minor groove, or along the backbone (Jones, 1999). Some proteins, including σ^{70} and cAMP-activated protein (CAP), interact with multiple surfaces of the DNA (Benoff *et al.*, 2002; Murakami *et al.*, 2002).

The most common type of DNA-regulatory protein interaction in prokaryotes involves a homodimer binding to a palindromic DNA sequence, with each half of the homodimer interacting with half of the palindrome (Huffman and Brennan, 2002). The majority of solved prokaryotic regulatory proteins to date bind DNA via a helix-turn-helix (HTH) motif, which is comprised of two alpha helices separated by a spacer with a length of 1 to 21 amino acids (Brennan and Matthews, 1989; Harrison, 1991; Wintjens

and Rooman, 1996). Several 'families' of HTH proteins have been identified from distinctive patterns of binding, but the common theme is that the majority of DNA-protein contacts are realized by a single alpha-helix inserted into the major groove of DNA, which yields sequence-specific contacts and hydrogen bonds (Wintjens and Rooman, 1996). This motif is found in prokaryotic regulatory proteins such as the *lac* and *trp* operon repressors and CAP, and in eukaryotic transcription factors such as *Drosophila hsf* and human *c-myb* (Chuprina *et al.*, 1993; Lawson and Sigler, 1988; Vuister *et al.*, 1994; Ogata *et al.*, 1994). The AraC family of regulatory proteins in *E. coli* is an exception to the homodimer rule: these proteins bind as monomers (Gallegos *et al.*, 1997). Crystal analysis of *Thermus aquaticus* RNA polymerase holoenzyme has shown that the σ^A subunit binds the –10 and –35 regions *via* HTH motifs, and makes non-specific contacts to the DNA backbone in the intervening sequence (Murakami et al, 2002). The winged-helix (WH) subset of HTH proteins, of which CAP is a member, supplement the familiar alpha-helices with a pair of alpha-helix 'wings' that make hydrogen bond contacts in the minor groove of a cognate DNA sequence (Gajiwala and Burley, 2000).

Several families of DNA-binding transcription factors are known in eukaryotes. Basic leucine zipper (bZIP) motifs are found in many eukaryotic transcription factors, including yeast Gcn4 (Konig and Richmond, 1993; O'Shea *et al.*, 1991) and human c-Fos and c-Jun (Glover and Harrison, 1995). The bZIP motif consists of a coiled-coil of basic alpha helices which interact with the DNA in the major groove (Efimov, 1999). Several different zinc finger motifs are used in eukaryotic DNA-binding proteins, of which the Cys(2)-His(2) type is most common (Wolfe *et al.*, 2000). The active structure of these proteins results from interactions between a cysteine-containing beta sheet, a histidine-containing alpha helix and a zinc ion. Transcription factors such as yeast Swi5 often contain multiple zinc-finger motifs in tandem, each binding independently to the major groove of DNA (Neuhaus *et al.*, 1992). The basic helix-loop-helix (bHLH) domain allows dimerization *via* amphipathic protein helices, which results in a bundle of basic residues that can interact with the major groove of DNA (Robinson and Lopes, 2000). Heterodimeric bHLH proteins include Max/Myc and Max/Mad (Amati and Land, 1994). The homeodomain is a DNA-binding motif found in many transcription factors required

for development and cell differentiation such as the mammalian *HOX* genes (Gehring, 1987). This motif is composed of three alpha helices, separated by a loop and a turn. As with the HTH motif, the last helix binds with functional groups and phosphates in the major groove of DNA (Qian *et al.*, 1989; Qian *et al.*, 1994).

While the above motifs all bind the major groove of DNA, some proteins bind the DNA in other ways. In particular, the key component of archaeal and eukaryotic promoter binding, TBP, binds primarily through hydrophobic interactions with the minor groove of DNA (Kim *et al.*, 1993*a*; Kim and Burley, 1994; Juo *et al.*, 1996). The critical amino acids are located in a beta sheet, while a series of alpha helices allow interactions with other proteins such as the eukaryal TAFs and archaeal TFB (Rashidzadeh *et al.*, 2003). Interactions between TBP induce a 90° bend in the TATA box, showing the important role of DNA deformability (Grove *et al.*, 1998). The bending of promoter DNA is a consistent feature in all three domains of life: bacterial σ has been shown to introduce a bend of 60 degrees in the promoter (Kunhke *et al.*, 1989).

Protein-DNA interactions

Specific interactions between amino acids and nucleotides have been extensively studied in the past decade. Protein mutagenesis experiments have been used to identify which residues are essential to binding (Griffith and Wolf, 2002) and hypotheses have been tested through the design of novel DNA-binding proteins (Corbi *et al.,* 1997; Sato *et al.,* 2002). With the structure of over 100 DNA-protein complexes solved, it is clear that there is no one-to-one 'code' of recognition between specific amino acids and specific nucleotides (Pabo and Nekludova, 2000). Instead, the interaction models developed thus far are specific to certain categories of binding protein such as the Cys(2)-His(2) zinc finger, and attempt to show the dependence of sequence recognition on the identity of specific residues. Functional proteins have been constructed and tested with these models, but the binding affinity of these engineered proteins is often suboptimal (Wolfe *et al.*, 2000).

DNA structure and flexibility have been implicated in the binding affinity of regulatory proteins. Werel *et al.* (1991) showed that deletion of a single nucleoside in the

σ^{70} melting domain yielded increased flexibility, and an increased affinity of the RNAP holoenzyme. The importance of DNA curvature, and the bending influence of RNAP and other binding proteins, have been extensively studied (Perez-Martin and Espinoza, 1994). Leger *et al.* (1998) showed that mechanical stretching of DNA increased the binding affinity of the RecA protein and suggested that this effect may be important *in vivo*. A study of eukaryotic Pol I promoters from many species (Marilley and Pasero, 1996) showed that while promoter sequence was highly variable between divergent species, the curvature and duplex stability was highly conserved throughout all eukaryotes. El Hassan and Calladine (1998) identified two different modes of DNA bending in response to protein interactions: either minor changes in helical path through deformation of flexible pyrimidine-purine hinges with drastic changes in major groove width, or a 'skewering' of dinucleotide pairs that dramatically decreases Twist and increases Roll without changing the width of the major groove.

Typically, there is a specific DNA sequence that the protein will interact with more strongly than all others, but strict adherence to the optimal sequence, required for some DNA-binding proteins, is not reqired for transcriptional regulatory proteins. The weaker interactions that occur with sub-optimal binding sites are often an important part of the regulatory strategy. For instance, the most common form of *E. coli* K12 RNA polymerase interacts most strongly with the consensus sequence TTGACA(N_{17})TATAAT, but this exact sequence is not found in the *E. coli* K12 genome; instead, each binding site is a variant with base substitutions, or a spacer of a different length, or both (Record *et al.*, 1996). While the spacer region shows little sequence conservation, the overall structure of the spacer is important in the helical alignment of the two conserved motifs (Auble and DeHaseth, 1988). An increase in the average number of G and C residues within the spacer decreased the rate of transcription initiation.

Juo *et al.* (1996) studied the nature of protein-DNA interactions at the promoter for TBP, and highlighted the impact of base substitutions that change the affinity of the binding protein. Their findings, when combined with the crystal data of TBP from *Arabidopsis thaliana* (Kim and Burley, 1994, Kim *et al.*, 1993*a*) and *S. cerevisiae* (Kim *et al.*, 1993*b*), led to important conclusions about the sequence and structural role of the

TATA box. The deformability of the TA step is likely instrumental in permitting the large deformations seen in the TBP/TATA box complex. While CG serves as a flexible hinge in some sequences that interact with major groove binding proteins (Schumacher *et al.,* 1994; Lewis *et al.*, 1996) another reason for preference of A and T is the steric hindrance of the guanine amino group, which projects into the minor groove and blocks TBP binding. In general, TBP has higher affinity for homogeneous T/A sequences than for sequences containing G/C pairs, and the introduction of a single cytosine or guanine at some TATA box positions can reduce TBP affinity by more than a factor of 10 (Juo *et al.*, 1996).

Detection of Regulatory Sequences

Why?

Information about regulatory sequences is generally derived through mutagenesis and footprinting experiments (Baliga, 2001; Munn and Alberts, 1991). While these experiments yield valuable information about important conserved sequences and regulatory proteins, they are time-consuming and cannot be carried out for an entire genome without an enormous effort. Such experiments have been carried out for model organisms such as *E. coli* K12 (Vo *et al.*, 2003), *Haloferax volcanii* (Thompson and Daniels, 1998), and *S. cerevisiae* (Teng *et al.*, 2001). These experiments have highlighted broad information about the mode of regulation in each domain, as well as specific details about the regulatory proteins in these model organisms. With computational methods such as BLASTP to identify similar regulatory proteins in related organisms (Altschul and Koonin, 1997), and phylogenetic footprinting to search for conserved binding sites (Gelfand, 1999; Gelfand *et al.,* 2000; Hardison, 2000), the experimental data have been successfully exploited to yield information from similar genomes.

However, since 1995 when the complete genome sequence of *Haemophilus influenzae* was determined (Fleischmann *et al.*, 1995), over 130 genomes have been sequenced and the data made publicly available, with the number of genomes steadily increasing. While many of these genomes simply represent variant strains of well-studied

species, many have been sequenced to increase the phylogenetic diversity of the data set, and have no close relatives for which experimental data are readily available. To learn about regulation in these genomes, computational methods must be devised to search for patterns in genomic DNA that may correspond to regulatory features. Also, since experimentally derived information is also biased in favour of certain genes and pathways of interest, these computational methods could also be used to extract additional regulatory information from well-studied genomes.

Challenges

Discovery of these conserved patterns in sequenced genomes is a difficult problem for several reasons. While coding sequence prediction methods can identify coding regions and translation start sites (Besemer *et al.*, 2001), less information is available for transcription start site predictions. Since the spacing between the transcription and translation start sites is highly variable in many genomes (O'Donnell and Janssen, 2001), the entire upstream region of a gene must typically be searched for regulatory patterns. This lack of precision requires the inclusion of a great deal of non-regulatory upstream sequence in the data set, which increases the level of noise that must be dealt with by a predictive method.

The key challenge in pattern detection is the variability in binding sites for a single regulatory protein. As mentioned above, no promoter in *E. coli* K12 matches the σ^{70} consensus perfectly, and the characterized variants are diverse in both sequence and spacing. In some cases, the promoter is far from the consensus and therefore very weak, and an activator protein is required for initiation of transcription (Rhodius and Busby, 1998). While many variants of the consensus sequence may occur, there may be covariation among sites such that a change in one position of the binding site may constrain the permitted changes at another site. A detection method must therefore consider the possible interactions among nucleotides at different positions in the sequence.

The diversity of regulatory proteins and their corresponding binding sites presents another problem for detection. Though there are thought to be >1000 transcription units

in *E. coli* K12 (Salgado *et al.*, 2000), some regulatory proteins may bind only a few of the corresponding upstream regions, and they may therefore be impossible to extract their binding sites from a large set of upstream sequences using statistical methods. Many of the characterized *E. coli* K12 repressors and activators have only a few known targets (Perez-Rueda and Collado-Vides, 2000). Different combinations of regulatory protein and sigma factor binding sites are also found in upstream sequences, an example is the *rpoH* gene in *E. coli* K12, which has at least three upstream promoters: three recognized by σ^{70} and one by σ^{E} (Gross, 1996).

Overview of methods

The problem of pattern detection in biological sequences has been actively studied for over twenty-five years. The functional and evolutionary relationships between homologous protein sequences led to the idea of sequence alignment. While sequence alignment methods are still being refined (Notredame *et al.,* 2000; Altschul and Koonin, 1998) to detect remote homologies, these methods already perform well on many sequences. The definition of the data set is of vital importance with pattern recognition and alignment methods: while coding sequences (and the encoded amino acid sequences) can be predicted using computational methods and validated with cDNA sequencing, it is more difficult to determine the correct boundaries of upstream regulatory regions.

Each of the following detection methods define the sequence data set for regulatory sequence detection in one of a few ways. Some are based on a subset of experimentally characterized promoters or other binding sites, which may or may not be aligned. The early promoter collections of Hawley and McClure (1983) and Bucher and Trifonov (1986), and the later collection of Ozoline *et al.* (1997), as well as sequences in the Eukaryotic Promoter Database (EPD: Perier *et al.*, 2000) and the Transcription Factor Database (TFD: Ghosh, 1992) are popular sources of data. A detection method is most likely to succeed on these data sets because all sequences are known to contain functional binding sites, and the exact location of binding sites are also usually known. If extensive non-coding sequence from a genome is available alongside gene expression data, then the upstream sequences of co-regulated genes can be searched for patterns. This approach

has gained in popularity with the advent of microarray analysis of gene expression (Britton *et al.*, 2002). Finally, some studies consider the whole set of upstream sequences from a genome: the goal here is to identify putative regulatory motifs that may be shared among a subset of genes, and to discover possible regulons. While this approach to data set definition requires the least amount of prior knowledge about regulation and can potentially reveal the most information, the data sets thus defined are prone to a high level of noise, due to the likely inclusion of DNA sequences that do not contain any regulatory motifs, and to the need to search for small motifs in large stretches of upstream sequence.

Many algorithms have been applied to the problem of regulatory sequence detection. While the methods described below are grouped by the type of algorithm used, different definitions and representations of the data set allow the same algorithm to be applied in many different ways. Different algorithms also share different characteristics: for instance, the statistical methods presented below, such as discriminant function analysis, are *exact* methods that invariably yield the same answer when trained with the same data set. In contrast, *heuristic* methods such as artificial neural networks and hidden Markov models use iterative methods to approach a solution, and replicated runs with the same data set will yield different answers that typically depend on a set of randomly chosen starting conditions.

Alignments

A common way of identifying conserved patterns in biological sequence is through sequence alignment. The goal of sequence alignment is to identify regions of similarity between different sequences, and line them up to yield an overall sequence model or to determine similarity and infer homology. Exact alignment methods such as Needleman-Wunsch (Needleman and Wunsch, 1970) and Smith-Waterman (Smith and Waterman, 1981) yield an optimal solution, but are too computationally intensive to apply to large numbers of sequences. Multiple sequence alignment of coding DNA or proteins is typically performed with a heuristic algorithm such as CLUSTALW

(Thompson *et al.*, 1994) or T-COFFEE (Notredame *et al.*, 2000), which yield global sequence alignments.

Two statistical alignment methods, Expectation Maximization (EM) (Lawrence and Reilly, 1990) and Gibbs sampling (Lawrence *et al.*, 1993), have been used to align conserved protein-binding sites in DNA. Both methods attempt to discover a motif within a set of sequences by constructing an optimal local alignment. Both algorithms construct an initial alignment, either randomly or with a 'guess' based on nucleotide frequencies, and will define the motif as occurring somewhere in the overlapping region of the alignment. The initial motif quality will likely be low, but is progressively improved by realigning the sequences and redefining the motif based on the new alignment. EM constructs a new alignment by searching each sequence for the best match to the defined motif, then realigning all sequences based on these best matches. The Gibbs sampler uses a 'leave-one-out' strategy, estimating the motif composition from the alignment of all but one of the sequences, then aligning the last sequence with the rest and choosing another sequence to leave out for the next calculation of the motif. Applications of the EM method include Multiple EM for Motif Elicitation (MEME: Bailey and Elkan, 1995), which allows multiple motifs and motif sizes to be searched, and MetaMEME (Grundy *et al.*, 1997), an EM method that optimizes a hidden Markov model to yield an optimal alignment. A variation on the Gibbs sampler was used to construct the BLOCKS database of conserved amino acid patterns in related proteins (Pietrokovski *et al.*, 1996).

EM was used in early studies to align conserved bacterial sequences. Lawrence and Reilly (1990) used an EM method to characterize the CAP binding site in 18 *E. coli* K12 sequences known to contain it, and successfully aligned and identified 16 of the sites. Cardon and Stormo (1992) implemented a modified EM algorithm that could model conserved sequences separated by a spacer of variable size, and used this method to align 231 *E. coli* K12 sequences from the Harley and Reynolds (1987) unaligned promoter set. Using proximity of the detected site to the known transcription initiation site as the criterion for success, this method correctly identified and aligned 87% of the promoter sequences. Each aligned pattern also had an associated probability score between 0.0 and 1.0 that signified its match to the consensus motif; however, no attempt was made to relate this quantity to promoter strength.

Gibbs sampling has also been used to detect motifs in DNA. The ANN-SPEC program (Workman and Stormo 2000) uses a Gibbs sampling method to optimize a motif model from a set of unaligned sequences, and models the motif using a single-output perceptron (see *Heuristic methods* below and Chapter 2). Recent Gibbs analyses have focused on generating a sophisticated background model to improve the detection of motifs. Thijs *et al.* (2001) used a high-order Markov model to represent background sequences, and were able to detect subtle introduced patterns inserted into random DNA sequences, and identify known binding sites from the upstream regions of co-regulated genes in *Arabidopsis thaliana*. (Liu *et al.*, 2001).

Text String-Based Methods

Text string-based methods have been extensively used to detect important patterns in DNA sequence. The goal of such methods is to detect 'words' of pre-specified or unknown length in biological DNA sequences that are present more or less often than would be expected. These words can be specific nucleotide strings with no substitutions allowed, or they can permit a certain degree of substitutions over the whole string, or degenerate characters at certain sites. However, the permissible amount of flexibility must be carefully defined, as too much specificity will cause many important patterns to be missed, while too much flexibility will yield excessive false positive results. The definition of the 'expected' number of words of a given type is also critical: this value can be obtained from a 'negative set' of sequences known to lack the site of interest, or can be derived from a background model that considers the frequency of oligonucleotides in the training set sequences.

Algorithms to detect over- and under-represented strings in DNA sequence include WORDUP (Pesole *et al.*, 1992), which uses a Chi-square test to identify nucleotide 'words' that are present to a greater or lesser degree than expected from a background model of mono- and dinucleotide frequencies. Other such methods use Z-score calculations, which differ in the background sequence model (Leung *et al.*, 1996; Schbath, 1997).

Hutchinson (1996) characterized promoters using frequency analysis of hexamers from these regions as well as coding and intergenic sequence. The most overrepresented hexamers in promoter sequences were easily associated with known features, including the CCAAT box and several variants of the TATA box. A similar study was performed on the upstream regions of co-expressed yeast genes (van Helden *et al.,* 1998), which detected known regulatory signals and other statistically significant patterns. A method that searched for constrained dyads with unspecified spacer regions (van Helden *et al.*, 2000) was later applied to detect sites bound by transcription factor dimers. An approach based on overrepresented strings was incorporated into an analysis by Zhang (1998), who used 30 nt windows to model different promoter features, and used linear and quadratic discriminant function analysis (DFA) to identify the promoter signals. Bussemaker *et al.* (2001) developed a model to correlate changes in gene expression levels with motifs found in upstream regions, which successfully identified elements involved in the induction of sporulation. The impact of these elements was assessed using a X^2 test that assesses the significance of their contribution to the change in expression levels.

An analysis described in Ohler *et al.* (1999) and used in Ohler (2000) uses interpolated Markov distances (IMD) to identify putative eukaryotic pol II promoters. The Markov chain describes the frequency of words of a given length N relative to the frequency of their constituent words of length $N-1$. Interpolation, which estimates the frequency of words of length N from shorter words, was performed to address the limitations of training set size, where small training sets would otherwise yield inaccurate parameter estimates. This model was highly effective at discriminating promoter regions from coding sequence, but was far less effective in separating promoter sequences from intron regions due to the closer similarity in sequence composition.

Crowley *et al.* (1997) and Crowley (2001) used Markov modeling to identify likely regulatory regions in DNA. The underlying principle of these studies is that eukaryotic non-promoter regulatory regions contain multiple repeats of the same binding site. Their system was trained to classify DNA as either regulatory or nonregulatory, and prior expectations of the number of regulatory regions were used to create the model. This analysis successfully identified known regulatory elements in eukaryotic DNA, with low background signal.

The MobyDick algorithm (Bussemaker *et al.*, 2000) represents DNA as a concatenated 'dictionary' of oligonucleotide words of length 1 or greater. The dictionary was 'grown' on a set of yeast DNA by extending constituent words if the extended versions were statistically significant. This approach yielded significant strings and pairs of strings separated by gaps, many of which corresponded to known binding sites.

Vanet *et al.* (2000) searched for conserved upstream patterns in the complete genomes of *E. coli* K12, *B. subtilis* 168 and *Helicobacter pylori* 26695. Their method searched for conserved words that were permitted a small number of mismatches, and were required to be in a minimum number of subsequences. Searches conducted for individual and paired motifs were successful in identifying canonical consensus patterns from *H. pylori* and *B. subtilis*, though a statistically significant result was not obtained for *E. coli*.

Kielbasa *et al.* (2001) implemented a system that searches for conserved regular expression-style strings using either a specific (A, C, G, T) or a degenerate (A, C, G, T, R, Y, S, W, K, M) alphabet, and scored the relative position of conserved strings in human regulatory regions. PromoterInspector (Scherf *et al.*, 2000) defines promoters in terms of their 'genomic context' as determined by a set of IUPAC strings. These strings are defined as a pair of substrings that can be separated by a 'wildcard' gap represented by one or more N nucleotides. Training of PromoterInspector involved the determination of which strings were overrepresented in promoter regions and which were more frequent in non-promoter regions. The functional assignment of a region of DNA sequence was then based on the relative abundance of these two types of words. This method yielded good prediction, with much lower false positive rates than previously observed with other methods.

The consideration of structure in building models of binding sites has been limited. Karas *et al.* (1996) built a structural profile of the eukaryotic TATA box based on minor groove width, and showed that different sequences can produce similar parameter values. This analysis did not provide optimal detection of a test set of TATA boxes, but the authors showed that the combination of sequence and structural analysis can yield improved detection. Ponomarenko *et al.* (1997) used helical conformation parameters derived from crystallographic data to describe the TATA box sequences of

yeast, invertebrates and vertebrates, and determined that sequences flanking the TATA box tend to have lesser twisting angles. While the activity of prokaryotic promoters tends to correlate with consensus similarity, weight matrix studies of eukaryotic binding sites did not show a similar correlation between matrix score and binding affinity. The ACTIVITY database (Ponomarenko *et al.,* 1999*a*) was designed to address this problem, and uses linear combinations of sequence and structural features to predict the activity of binding sites.

Weight matrix-based methods

A weight matrix represents a motif by assigning a score for each base at each motif position. A DNA sequence of the same length as the weight matrix can be scored with the weight matrix by summing the scores for matching bases at each position. This summed score can then be compared with a threshold value to determine whether the DNA sequence matches the motif modeled by the weight matrix (Stormo, 2000). Weight matrices are often preferable to consensus sequences in modeling and detection of motifs, because they can model the *information content* of each position in the binding site. Nucleotide bases at highly conserved positions within the motif will have a large impact on the score, while weakly conserved positions will make very small contributions to the matrix score.

The CONSENSUS program (Stormo and Hartzell, 1989; Hertz *et al.,* 1990) and its derivative WCONSENSUS (Hertz and Stormo, 1995) build frequency matrices from unaligned sequences in a progressive manner. All k-words (DNA subsequences of a given length) from the first sequence in a training set are used to generate simple weight matrices that represent the exact base at each position with a value of 1.0. Thus, the short sequence AGTGCC could be used to generate weight matrices that match the 3-words AGT, GTG, TGC, and GCC exactly. These frequency matrices are updated by adding each training set sequence in turn, and using the k-words to update the existing matrices. After each new sequence is added, the updated matrices are screened using either a cutoff of information content, or by keeping only a predetermined number of the best matrices. The advantage of this method over exact multiple sequence alignment is that each

sequence need only be compared with the derived frequency matrices, and not with every other sequence in the training set. This heuristic approach yields a dramatic reduction in computational time. Statistical determination of weight matrix significance can also be performed (Hertz and Stormo, 1999; Claverie, 1994).

The MatInd program (Quandt *et al.,* 1995) takes as input a frequency matrix derived from a set of known binding sites, and generates a vector that represents the degree of conservation at each position within the site consensus. MatInd also defines a 'core region', comprising the most highly conserved set of four consecutive nucleotides, for use in signal detection using MatInspector (see below). MatInd is implemented by the eukaryotic transcription factor database (Wingender *et al.*, 1996). DMS (Hu *et al.*, 2000) uses a more complex system to build weight matrices from a set of related sequences, considering the quality, specificity and frequency in a set of related sequence. These matrices are then used as the basis of a decision tree approach to classifying sequences. Li *et al.* (2002) used a whole-genome approach to building weight matrices by looking for overrepresented pentamer sequences separated by a gap of a variable length, then clustering pentamer pairs with consistent spacing and using these clusters to construct weight matrices. Many of the weight matrices thus obtained from the *E. coli* K12 genome were very similar to weight matrices constructed directly from known binding sites.

Ponomarenko *et al.* (1999*b*) introduced several new nucleotide alphabets for use in weight matrix representations. In addition to the traditional position-specific representations, weight matrices were constructed using alternative representations such as dinucleotides, different groupings of nucleotides, and nucleotides with unknown spacers. These separate alphabets were all used to model the NF-1 transcription factor binding site, and some yielded improved prediction accuracy on a set of known sites.

Many algorithms have been developed to search for the patterns defined by weight matrices in DNA sequences. MatInspector (Quandt *et al.*, 1995) uses the output from MatInd to search for modeled binding sites. The similarity of a sequence to the matrix is dependent on the match of each base at each position to the frequency matrix, as well as the degree of conservation (the 'consensus score') at each site. A low consensus score at a site minimizes the contribution of that site to the overall matrix score.

Similarly, MatrixSearch (Chen et al, 1995) searches a set of sequences for matches to weight matrices contained in its set, the Information Matrix Database.

Since individual DNA-binding sites are short, and found in non-regulatory as well as regulatory regions of DNA (Prestridge and Burks, 1993), several methods detect regulatory regions by searching for specific element combinations. GenomeInspector (Quandt *et al.*, 1996*a*,*b*) searches for distance relationships between the binding sites of different regulatory proteins within upstream regions. The ModelGenerator and ModelInspector programs (Frech *et al.*, 1997) combine this positional information with the detection of other sequence elements such as short repeats to build complex models of upstream regions. These programs were used to create a model of human and avian retroviral promoter elements, and this model was used to detect previously unknown functional promoters in these retroviral genomes. The FastM program (Klingenhoff *et al.*, 1999) is another method that analyzes the variable spacing within composite regulatory elements, and generates models that can be tested with ModelInspector. However, while ModelGenerator requires a training set of size >10, FastM can use even a single sequence to build its model, since it relies on the user to define model parameters. PROMOTER SCAN (Prestridge, 1995) combines string searches for eukaryotic transcription factor binding sites with a weight matrix-based search for the TATA box.

Thieffry et al. (1998) used a weight matrix and string search approach to detect putative regulatory sites in the recently sequenced E. coli K12 genome. The weight matrices and strings were derived from known binding sites of 56 regulatory proteins, and were used to search 400 nt upstream and 50 nt downstream of the start codon of each annotated coding sequence. Many plausible new targets for known regulatory proteins were found, though the upstream regions of most annotated coding regions had no predicted regulatory sites. The authors pointed to our limited knowledge of the number of regulatory proteins in even a well-studied organism such as *E. coli* and suggested that some of the missed genes would be targets of as yet uncharacterized regulatory proteins.

Heuristic methods – Hidden Markov Models and Artificial Neural Networks

Hidden Markov models (HMMs) model a conserved sequence as a series of states, which can exist either as matches to the consensus pattern, or insertions and deletions that affect the overall length of a pattern (Baldi and Brunak, 2001; Sonnhammer *et al.*, 1998). The key to pattern prediction with an HMM are the probabilities associated with a match for each position, as well as the transition probabilities from one state to the next. These probabilities must be trained with a set of known sequences. HMMs have been applied to biological problems such as translation initiation site prediction (Yada *et al.*, 1997) and sequence alignment (Madera and Gough, 2002) as well as promoter recognition.

HMM learning of prokaryotic and eukaryotic promoter sequences was demonstrated by Pedersen *et al.* (1996). The promoters used to train the HMMs were all aligned with the transcription start site, and separate HMMs were trained with 166 *E. coli* and 90 human promoters. Analysis of the information content of the trained HMMs showed that the canonical features were accurately modeled. Yada *et al.* (1997) built HMM representations of eight σ factors in *Bacillus subtilis*, using known promoters and similar promoters from other genomes to increase the size of the training set. The trained HMMs were good classifiers of the training set, and a genomic analysis of *B. subtilis* showed good correspondence between the predicted σ factor and the role of the gene.

Artificial neural network (ANN) methods have been used to detect several different types of biological pattern. Protein features detected and classified with ANNs include transmembrane domains (Lohmann *et al.*, 1994) and cleavage sites (Narayanan *et al.*, 2002; Schneider and Wrede, 1994). In addition to promoters, other DNA features identified with ANNs include transcription start sites (Pedersen and Engelbrecht, 1995), ribosome binding sites (Bisant and Maizel, 1995), and intron/exon compositional differences (Granjeon and Tarroux, 1995).

Several analyses in the early 1990s used the backpropagation algorithm (see Chapter 2) to detect promoter sites in DNA. O'Neill (1991, 1992) used aligned promoter

sequences as the positive set to train a backpropagation neural network, with random sequences of 60% (A+T) composition and known promoter down-mutations as the negative set. The positive set in this experiment was enlarged by introducing base substitutions at non-conserved sites in real promoters, then adding these new sequences to the positive set. The trained promoter achieved a generalization accuracy of 80% on the test set of promoters, with a false positive rate of less than 0.1%. Rules extracted from the network clearly showed the vital role of the –10 and –35 sequences, but implicated bases flanking these regions as well. Demeler and Zhou (1991) trained a backpropagation neural network to distinguish between characterized promoters and random DNA sequences, with separate runs for the –35 and –10 regions as well as a combined run. The trained networks were tested on the genome sequence of phage fd, and 9 of 11 known promoters were identified. However, at least 160 false positives were obtained in the 6408 base pair sequence.

NNPP (Reese, 2001) uses neural networks to detect conserved sequence features in eukaryotic promoters. This method implements a 'time-delay' architecture, which imposes constraints on the connections between input and hidden layer nodes (see Chapter 2) that allow position-independent detection of various signals. This architecture allows the system to detect features such as the TATA box and the initiator region that have inconsistent spacing. This method compared favourably to previous methods when run on a test set of known promoters. ANN-SPEC (Workman and Stormo, 2000) trains perceptron networks to recognize short, specific sequences. This approach was combined with CpG island detection in a later study (Hannenhalli and Levy, 2001).

A novel approach to neural network-based pattern detection was carried out by Knudsen (1999). The program Promoter2.0 creates a random population of neural networks, and chooses those that yield the best discrimination between a positive set containing known promoters, and a negative set lacking promoters. The neural networks thus tested also used the predictions of previously tested neural networks as input; this allowed the representation of multiple features such as the TATA-box, BRE, and CCAAT sequences. After each pass, neural network weights were modified to try and improve generalization. Using the Bucher and Trifonov (1986) collection of eukaryotic promoters, correlation coefficients on a test set of sequences reached 0.63. The trained

neural networks were assessed by presenting them with all 4096 hexamers, and the best scoring hexamers for each neural network could clearly be associated with promoter features.

The Dragon Promoter Finder (DPF) (Bajic *et al.*, 2002) combines a weight matrix approach with neural network analysis to yield promoter predictions. The DNA sequences presented to DPF are first interpreted using a set of weight matrices that can distinguish between intron, exon, and promoter-containing sequences to yield predictions of sequence context. These predictions are then presented to an artificial neural network, which uses information about these different regions to predict the transcription start site and close promoters.

My Approach, in Brief

Since DNA structure and flexibility are crucial determinants of DNA-protein interactions, an informative model of a regulatory element or promoter should be able to incorporate conformational parameters in its representation. This type of information can be useful in increasing the specificity of a model, because a single, specific structural constraint can replace multiple oligonucleotides. The goal of my research is to develop algorithms that can detect patterns in DNA using information about DNA sequence and sequence-dependent DNA structure.

Instead of choosing a single method to represent motifs in DNA, the algorithms described in the following chapters can use any of several different ways to build motif representations. Patterns can be described with IUPAC strings, as scoring matrices, by their common structural properties, or as combinations of all three. The flexibility of these representations requires an effective method to search through model space: the methods described in this thesis use a genetic algorithm-based approach to construct and test sets of models with the goal of converging on one or more optimal models.

Another important principle of these methods is the ability to construct models without any prior knowledge of binding sites. Since many sequenced genomes are otherwise poorly understood, identification of regulatory features will depend in large part on the detection of novel features in genomic sequence. However, these methods can

also use prior knowledge to help with model construction, in the form of predefined strings or weight matrices, or more general knowledge of motif length and composition.

This thesis is concerned with the automated detection of regulatory protein binding sites and promoters. My primary goal is to address this with only two pieces of information as a starting point: (1) A complete genome sequence, and (2) A table of predicted open reading frames (ORFs) within that genome. While focusing on this difficult problem, it is worth remembering that a wealth of experimental evidence *is* available for many genomes, most notably the model bacteria *E. coli*, *B. subtilis*, and model eukaryotes such as *S. cerevisiae* and *H. sapiens*.

This thesis is organized into seven main sections: this Introduction, five chapters describing methods and experiments, and a general Discussion. The Genetic Algorithm Neural Network (GANN) software, which underlies many of the experiments presented here, is presented and described in detail in Chapter 2. Chapters 3 and 4 all use this software, but each of these chapters contains a Methods section where specific implementation details of GANN are described, as well as the strategy used to represent DNA. Chapter 5 describes the Motif Genetic Algorithm (MGA) software, which uses evolutionary methods to detect patterns in genomic DNA sequence. This approach is used to generate sequence and structural models for known binding sites in Chapter 6.

The experiments described in this thesis are based on three different strategies for representing DNA, and are summarized in Chapters 3, 4 and 6. Chapter 3 describes an attempt to find over-represented patterns in whole upstream regions of *Escherichia coli* K12 using an early version of the GANN software. The experiments in Chapter 4 use a sliding window method to subdivide the DNA, and introduce simple methods of converting DNA to neural network inputs. These methods are carried out in two separate genomes: those of *E. coli* K12 and the thermophilic crenarchaeote *Sulfolobus solfataricus* P2. An alternate statistical method, discriminant function analysis, is also performed on the data for comparative purposes. In Chapter 6 the MGA is used to generate sequence and structural models of regulatory protein binding sites from *E. coli* K12, using either characterized binding sites or upstream regions of co-regulated genes.

The general conclusion at the end proposes ways to improve GANN and MGA, and describes how other types of analyses could be performed using these software packages.

[2]

The Genetic Algorithm Neural Network (GANN)

Overview

There are numerous statistical methods and heuristic algorithms that can be used to identify patterns, and use these patterns for classification. Artificial neural networks (ANNs) have several properties that make them well-suited to the task of motif detection in DNA. There are also limitations and specific considerations that need to be addressed when using neural networks; these are described later in this chapter.

The key advantage of ANNs over other methods such as consensus sequences and weight matrices is their ability to model interactions between different characteristics of the input patterns (Baldi and Brunak, 2001). When presented to an ANN, a sequence of DNA or protein is typically modeled as a series of nucleotides or amino acids. The encoding used is typically 'sparse', with either 4 (for nucleotides) or 20 (for amino acids) inputs for each sequence position. This contrasts with dense representations, which can model single nucleotides with only 2 input nodes, since a pair of binary nodes can have any of 4 different states. Dense encoding requires fewer input nodes, but introduces false correlations between bases (Wu, 1997). A specific base at a given position will be represented by a nonzero value at one of the associated inputs and zeroes for the others. The input signal is passed through the ANN to yield numerical values at a set of output nodes, which represent the network's classification of the input pattern.

ANN Architectures

The connectivity and function of the nodes define the architecture of a neural network. The multitude of ANN architectures can be classified in several ways, but an important division can be drawn between two major types: those with recurrent connections between nodes, and feedforward nets without recurrent connections (Jordan,

1999). Recurrent connections within an ANN yield a cyclical graph configuration, and a prediction is obtained when the recurrent propagation of values reaches a stable result. An example of a recurrent network architecture is the Bidirectional Associative Memory (BAM) network (Kosko, 1988). A BAM network consists of a set of nodes organized into an input layer and an output layer. The input pattern is presented to the nodes in the input layer, and the node values and connection weights are used to determine the output layer node values in a manner similar to multi-layer perceptrons, presented below. The output node values are then passed back to the input layer via the same connections, and this cycle continues until the node values reach a stable state and are not changed by further propagation of values. This method, however, is limited in the number of patterns it can learn reliably (Rogers, 1996). Another type of recurrent network is the Adaptive Resonance Theory network, which classifies input patterns by comparison with model vectors, and introduces new model vectors when necessary (see **Learning Paradigms** below).

The non-recurrent, feedforward ANN architecture has been used to detect and classify many different types of patterns. Within the context of pattern detection in molecular sequence data, feedforward ANNs have been applied to problems such as protein family classification, secondary and tertiary structure prediction, and gene detection (Wu, 1997). The most popular architecture in biological pattern recognition is the multi-layer perceptron model (MLP).

The MLP architecture consists of a series of nodes organized into separate layers connected by interlayer links (Figure 2-1). Each node in the network is assigned a scalar value either directly from the input pattern, or from the incoming connections to the node. Each link between nodes in adjacent layers has an associated connection weight that determines the influence of its source node on the final value of its target node. An MLP has an input layer of nodes that receives the encoded patterns, and an output layer that represents the network's prediction. There may also be one or more hidden layers which are intermediate between the input and output layers. To make a prediction, input layer node values and the connection weights leading to the next layer are used to determine the node values in the next (hidden or output) layer. This process is repeated until the node values in the output layer have been calculated. Learning of patterns is performed

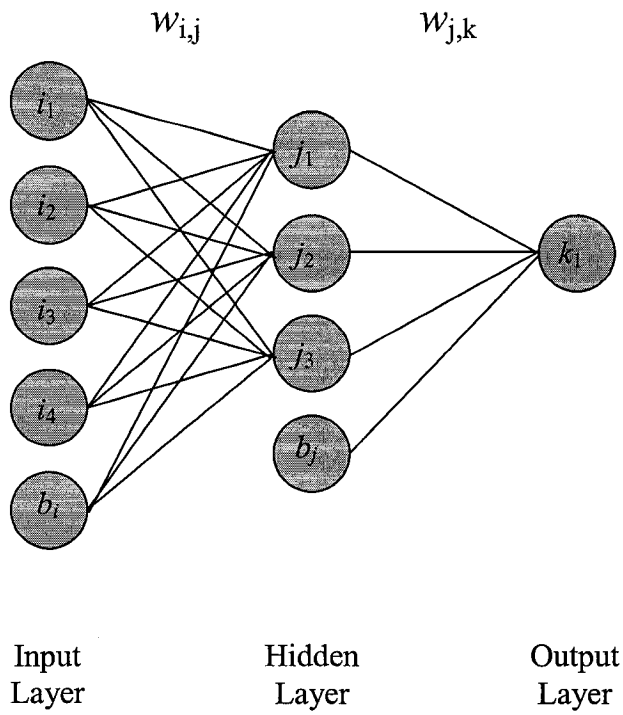


Figure 2-1: Architecture of a generic multi-layer perceptron (MLP). The nodes in adjacent layers are completely interconnected, except for the bias nodes which connect only to the nodes in the next layer forward.

by changing the connection weights to yield better predictions on a training set of data for which target values are known.

A simple version of the MLP is the two-layer perceptron network (Rosenblatt, 1962). The two-layer perceptron consists of a layer of input nodes, each connected to the single output node. Signals presented to the k input nodes ($x_1, x_2, \ldots, x_k$) propagate to the output node y *via* a summation function:

$$y = f(\Sigma\, x_i\, w_i) \qquad (2.1)$$

where w_i is the weight associated with the connection between input node x_i and the output node y. $f(\,)$ is a threshold or activation function used to constrain the output node values to a specific range.

The two-layer perceptron, with its input and output layers, can learn to solve some classification problems. This type of network has been applied to biological problems such as ribosome binding site prediction (Stormo *et al.,* 1982) and promoter recognition (Bajic *et al.,* 2002, Knudsen, 1999) and has yielded better classification accuracy than rule-based systems (Baldi and Brunak, 2001). This system is analogous to a weight matrix representation of DNA, where each base has an associated score at each position within a sequence. The combination of base scores yields a prediction, which can be used to predict the function of the sequence. This similarity has been applied in weight matrix training: Stormo *et al.* (1982) trained a two-layer network to detect *E. coli* ribosome binding sites (RBS), then interpreted the trained connection weights as a weight matrix representation of the RBS sequences.

A limitation of the two-layer feed-forward model is that it cannot accurately solve mapping relationships that are not linearly separable (Mehrotra *et al.*, 1997). If the relationship can be expressed as an inequality in multidimensional space, then a two-layer network may be able to learn the correct mapping. The historical example that illustrated this limitation is the 'exclusive OR' function, which has two inputs that can each take on a value of 0 or 1. The function returns 'true' if one of the inputs is 0 and the other is 1, or 'false' if the inputs both have the same value. When plotted in two dimensions, the two types of result cannot be separated from one another

with a single line. There is no set of connection weights for a two-layer feedforward network that can accurately model the XOR function, and the same is true of more complex functions that are not linearly separable (Bose and Liang, 1996).

The addition of hidden layers of nodes to the MLP architecture allows the classification of functions that are not linearly separable. The hidden layer nodes allow the modeling of interactions between input units (Bishop, 1995). Three-layer MLP networks were applied to the problem of *E. coli* promoter recognition in the early 1990s (O'Neill, 1991; O'Neill, 1992; Demeler and Zhou, 1992). When these networks were trained on a set of characterized promoters, prediction accuracies of close to 100% were achieved, but these predictions were typically specific to the length of the spacer between the conserved –35 and –10 regions, because promoter alignment was critical to the training of the ANN.

Some variants of the basic MLP architecture do not rely on the typical pattern of layers that are completely linked by independent connections. An example is the Time Delay Neural Network (Waibel *et al.*, 1989), which was initially developed for speech recognition. Instead of complete connections with the input layer, each hidden node has a more limited *receptive field* that links it to a contiguous set of input nodes. The connection weights between a hidden node and each input node in its receptive field are constrained to have the same value. The separation of input nodes into overlapping receptive fields allows any input node within a given field to represent the feature that is detected by that field, thus reducing the need for features in the input data to be correctly aligned. Reese (2001) used this model as the basis for the NNPP program, which trained two separate TDNNs on two windows of sequence: one containing the eukaryotic TATA box and the other containing the initiator element. This trained system detected 70% of promoters, with an estimated false positive prediction rate of 1/547 bases.

Another class of non-recurrent ANNs are the Radial Basis Function (RBF) networks (Bishop, 1995). As with other types of ANN, an input vector is presented to the input nodes of an RBF. However, instead of containing a single scalar value, the nodes in an RBF layer each present a vector of the same length as the input vector. The activity of each node depends on the distance between its vector and the input vector, hence the 'radial' nature of the activity, where more distant vectors will yield lower activation

values. These RBF nodes can then pass their results on to an output layer *via* weighted connections, or can compete with one another to determine which one best matches the input signal. An example of a competitive network is the Kohonen Self-Organizing Map (see **Learning Paradigms** below).

Learning Paradigms

There are many methods that have been applied to the problem of training ANN connection weights or nodes. The methods used to train classification systems such as ANNs can be divided into three distinct categories: unsupervised, supervised and reinforcement learning. The key difference between these schemes lies in the response of the system to the presented training patterns: whether the system is aware of the desired classification for the training cases, and whether the system responds directly to each training case.

Unsupervised Learning

The goal of unsupervised learning is to classify patterns based on their similarity to other patterns, without any outside specification of how these patterns should be assigned. ANNs that implement unsupervised learning methods are analogous to statistical clustering algorithms (Schalkoff, 1992), which also seek to classify data points into well-defined groups, without a prior definition of what those groups should be. The Adaptive Resonance Theory (ART) (Carpenter and Grossberg, 1988) network accepts a vector of data at its input nodes, and passes these values along connections to nodes that compare this vector with a series of model or prototype vectors that represent different classification groups. The output node that is most similar to the input vector will be stimulated; however, the output nodes then send their values back to the input layer through weighted feedback connections. This cycle continues until a specific output pattern is predicted, and this prediction does not change with further cycling of signals. The network is 'Adaptive' because it can add new output nodes to its architecture if presented with a pattern that is not sufficiently similar to any of the existing model

vectors. The key to this determination is the vigilance term, which represents the required similarity for an input vector to match a prototype pattern (Mehrotra *et al.*, 1997).

The Kohonen Self-Organizing Map (Kohonen, 1988) is a two-layer RBF network that compares an input vector to a multidimensional map of neurons, each containing a vector of the same size. The target vector that is most similar to the input vector will be stimulated, and this mapping represents the classification of the input pattern. The target vectors in a Kohonen network are typically given random initial assignments, and the target space is trained by presenting a set of training vectors to the network. The target node stimulated by a training vector is modified to increase its similarity to the input vector, and other target nodes in the neighbourhood of the stimulated node are modified to a lesser degree. The result of this training operation is a smoothed set of vectors in the target space, where similar input vectors will stimulate the same or neighbouring target neurons. While unsupervised learning has been applied to biological classification (Wang *et al.*, 2001; Cai *et al.*, 2001) and presents an intriguing way of classifying DNA sequences, these methods are not implemented by GANN.

Supervised Learning

In supervised learning, each training case is presented to a system in turn, and the system is modified directly to yield a more accurate prediction on the current training case (Reed and Marks, 1999). These methods rely on an estimate of the prediction error to determine the magnitude of the change to be applied to the system. The backpropagation algorithm, which uses prediction error to modify connection weights, is often used to train MLP networks. The error δ_k of an output layer node k is based on the difference between the predicted (p_k) and target (t_k) values for each output node applied to the derivative of the tanh function (Bishop, 1995):

$$\delta_k = 1 - (p_k - t_k)^2 \qquad (2.2)$$

Since hidden layer nodes are internal to the system and cannot be compared with any expected values, defining suitable error values for them is more difficult. (Rumelhart

et al., 1986) developed a method that defines the error δ_j of a hidden layer node j in terms of the node's contribution to the error of the output nodes:

$$\delta_j = 1 - (\sum w_{jk} * \delta_k)^2 \qquad (2.3)$$

where w_{jk} is the connection weight between hidden layer node j and output layer node k, and δ_k is the error associated with output node k.

Once error values for each output and hidden layer node have been determined, they are used to change the value of the connection weights in the ANN. The change in the weight of a connection between node a in the input or hidden layer, and a node b in the hidden or output layer, respectively, is as follows:

$$\Delta w_{a,b} = -\eta * \delta_b * x_a \qquad (2.4)$$

where δ_b is the error of node b, x_a is the value of node a, and η is a multiplier called the learning rate, which is either constant or a function of the number of elapsed training rounds.

A round of training progresses by presenting each member of the training set as input to the ANN in turn. Since the desired output is known for each member of the training set, the neural network predictions can be compared to the target output to determine the output node error, as above. One drawback of training an ANN with an algorithm such as backpropagation is the problem of *overfitting*, where the ANN 'memorizes' each pattern without learning the important features that underlie the training sets. Overfitting is most likely to occur with very large networks and small training sets: the large number of connection weights will likely be able to yield individual representations for each training set member (Kasabov, 1996). The situation is analogous to using an excessively large number of variables to generate a statistical model, which may yield excellent predictions on the training set because of optimization of the random variation across many variables. Overfitting can be assessed in several different ways, such as bootstrap or jackknife analysis, or data splitting. Data splitting produces a training set that is used to generate a model, and a test set that can be used to

assess the model's *generalization* ability. Poor prediction accuracy on the test set is suggestive of overfitting, while good performance on the test set indicates that the model has learned important rules about classifying the data. The validity of this conclusion depends on the independence of the training and test sets, so bootstrapped data splitting is often performed, where the random separation of known data into training and test sets is performed several times (Hassoun, 1995).

Another limitation of gradient-descent methods such as backpropagation is their tendency to find a solution that is locally optimal but does not represent the best overall solution (Hassoun, 1995). Small, incremental changes allow the model to search for improvements in a limited range of the error surface, but if a local minimum is reached the algorithm may get stuck and be unable to find a better set of parameters. Tools to escape local minima include using a *momentum term* (Rumelhart *et al.*, 1986) which remembers previous parameter changes, and uses these previous changes to increase the magnitude of the current change. A *global-descent* variant of backpropagation attempts to break out of local optima by remembering the set of connection weights that yielded the last local optimum, but includes a 'repeller' that tries to push the weights away from that optimum, hopefully to a better optimum (Zak, 1989).

Reinforcement Learning and Evolutionary Methods

Reinforcement learning is used to modify the behaviour of a system, without a specific response to individual training cases (Sutton and Barto, 1998). The system's *policy* is the method it uses to perform a specific task, such as pattern classification. The effectiveness of a given policy is evaluated using a *reward function,* which describes how well the system has performed its task. A *value function* can also be used to estimate the long-term reward potential of a policy, if the system is capable of making such projection. Evolutionary methods such as genetic algorithms (Mitchell, 1998), simulated annealing (Davis, 1987), and genetic programming (Koza, 1992) implement a form of learning similar to reinforcement learning, though they do not use value functions for fitness projections.

A genetic algorithm (GA) simulates biological evolution by evaluating the ability of multiple sets of parameters to perform a task. When applied to neural networks, the task to be optimized is the classification of patterns, either through the determination of optimal ANN architecture, or optimization of connection weights, or both (Mitchell, 1998). In a GA, a complete set of parameters (such as a set of connection weights) used to describe a system are arranged in a *GA Chromosome*, with each parameter represented by a single gene. An instance of a GA Chromosome will specify a value for each required parameter. A population of GA Chromosomes, each with different parameter values, can then be evaluated against a training set of data such as a set of encoded DNA sequences. A *fitness* is assigned to each GA Chromosome based on its accuracy in performing the required task on the training data, and the entire population can be ranked by this criterion. The relative fitness can be used to decide which members of the population are used to create a new generation of GA Chromosomes, subject to operations such as recombination and mutation (see **GANN** below for defined evolutionary operations).

When used to train parameters such as ANN connection weights, the GA training method is less likely to get trapped in local error minima than a gradient-descent method such as backpropagation (Hassoun, 1995). If parameter training is performed with the goal of reaching an optimally low error through small, gradual changes, then the optimum that is eventually reached may be dependent on the starting parameters and may not be optimal. The GA allows simultaneous searching of many different combinations of parameters, and evolutionary operations may allow parameters to escape local optima. The generation of random combinations of parameters, and the use of recombination and mutation to generate new combinations of good parameters, can yield more efficient optimization than a complete enumeration of all possible combinations of parameter values (Back, 2000).

GANN

Overview

GANN (Genetic Algorithm Neural Network) is a suite of programs designed to detect patterns in data using backpropagation- or genetic algorithm-trained neural

networks as the classifiers, and genetic algorithms to choose the variables and neural network architecture that yield the best prediction accuracy. This software was developed with the problem of regulatory sequence and promoter detection in mind, but can be applied to other classification problems as well.

GANN is accompanied by software to extract genomic sequence data and assign categories to extracted sequences based on their positional context: if regulatory sequence detection is the goal, then regions upstream of the start codon will constitute the positive set, while other intergenic sequences and (optionally) protein-coding regions are assigned to the negative set (see **Set Definitions** below).

A single experiment involving GANN will use either backpropagation or a genetic algorithm to optimize the connection weights of a three-layer MLP network. Optimization is based on a network's accuracy in classifying members of the positive and negative training sets. Since overfitting of any trained system is a concern, the neural networks are also presented with a test set that is not used during the training phase when connection weights are modified. The classification accuracy on the test set indicates the network's generalization ability. The Outer Genetic Algorithm (OGA) is used to choose among the available indices that represent sequence features, as well as the different options affecting neural network size and training parameters. The optimizations carried out by the OGA are based on the generalization ability of the networks it creates, with the ultimate goal of producing ANNs that are well-suited to learning the important rules for classification of sequence patterns.

GANN for Parameter Optimization

To present DNA sequence and structure to a neural network, the DNA is converted to a set of floating-point numbers to represent characteristics such as structural features and oligonucleotide counts. The sets of values derived from these conversions are here called *indices*. Since all indices rely on interpretations of the four-letter DNA sequence, the indices generated with different sequence and structural rules will likely not be entirely independent of one another. Useful information may be shared or even duplicated among indices, in which case a predictive model need not use all of these

indices. Also, some parts of the upstream DNA sequence may be devoid of any information of use for predictive purposes, either because there are truly no biologically relevant patterns, or because the patterns are so rare that they cannot be detected in the data set by statistical means. For these two reasons, a predictive model based on DNA-derived indices should not include all of these indices, but should instead be able to choose a subset of all indices that yield maximal information, and whose contributions are complementary to one another.

Another challenge lies in the optimization of neural network parameters. There are parameters that are common to both backpropagation-trained (BPNet) and genetic algorithm-trained (GANet) neural networks, such as hidden and output layer size, and parameters that are specific to the learning algorithm, including the learning rate and momentum term for BPNet, and the recombination and mutation rates for GANet. Since the parameters to be optimized can exert combinatorial influences on the model, they cannot each be optimized in isolation. A strategy that considers many combinations of parameters is necessary.

The GANN system is used to optimize the neural network architecture and parameters. As with the best indices, the parameters associated with networks that have high generalization scores will be preferentially used to define the next generation of neural networks.

GANN uses an evolutionary method to choose the indices that yield the best test set or *generalization* scores, and generate new combinations of these indices which may yield even higher scores. Indices that yield poor generalization scores will eventually be replaced by new copies of good indices, and become extinct. This approach therefore serves not only as a classification system, but as a method to identify and to select the indices that can best discriminate between the positive and negative sequence sets.

Each instance of a GANN parameter set (or OGA Chromosome, see **Design and Implementation** below) specifies both a complete generative parameter set for either BPNet or GANet, and a set of indices to use as ANN input, sampled from the entire set of indices. The following evolutionary operations are defined for GANN:

- RECOMBINATION: A pair of parameter sets (OGA Chromosomes) with high fitnesses each contribute a random subset of their parameters to create a new OGA Chromosome. Each parent randomly contributes exactly half of its inputs and neural network parameters to the offspring, with the constraint that the same input cannot be inherited from both parents. The new OGA Chromosome will then replace a low-scoring Chromosome from the current population. The fraction of parents to consider and the fraction to replace during each round of optimization are specified by the user.
- MUTATION: After recombination, the ANN parameters specified in an OGA Chromosome are subject to mutation. The fraction of the population with lowest fitness (which may include some new recombinants) is subject to mutation. The important user-specified variables represent the proportion of Chromosomes subject to mutation, the probability that a given parameter will mutate, and the maximum proportional change in the parameter value. No mutation is performed on the list of specified input indices.
- MIGRATION: If more than one interbreeding population of OGA Chromosomes exists, then a population may export a Chromosome to another population at a predefined frequency.

The neural network parameters specified by an OGA Chromosome are each constrained to a user-specified range of values. For instance, the number of hidden nodes in the ANN cannot be less than 1, and a maximum value is needed to prevent giant networks that execute slowly. If a parameter is pushed outside its legal range by a mutation event, the parameter value will be set to the corresponding limit at the end of the evolutionary phase. After the range checking and correction is complete, another round of OGA training and testing can be performed.

Each MLP network created by a specific OGA Chromosome is trained and tested for a predetermined number of rounds. While ANNs are often trained until a certain level of prediction error is reached, the use of a fixed number of training rounds ensures that each parameter set and architecture can be judged at the same point in training. While many ANN training runs are carried out until a target level of overall prediction error is

reached, choosing such a target would be inappropriate for GANN. Since many different parameter combinations are evaluated by GANN, a global error target would either be too low and never be reached by poor parameter combinations, or would be too high and fail to distinguish between good parameter combinations, which would all reach the target very quickly during training.

The parameters controlling recombination and mutation of OGA Chromosomes could not be optimized due to computational limitations. Therefore, the values chosen for the experiments described in Chapters 3 and 4 were chosen arbitrarily, but with specific goals in mind: a minority of OGA Chromosomes (30%) were allowed to recombine, and the recombinants thus obtained replaced the worst performing 30% of the population. Seventy percent of the population was susceptible to mutation, thus potentially affecting all recombinants and all OGA Chromosomes that did not recombine. When mutated, the change in an OGA Chromosome parameter would not exceed a factor of 1.1, which prevented rapid change in these parameters. The number of OGA Chromosomes was specific to each GANN experiment (see Chapters 3 and 4), and was chosen to ensure a desired degree of sampling of all input indices and combinations.

MLP Architecture

With the exception of the input vector size, which is predetermined, the important architectural features of the neural network are optimized by the OGA. The number of hidden and output nodes are assigned legal ranges as described above. Since the layers of the network are completely interconnected, the number of connections is determined by the number of input, hidden, and output nodes. A bias node, which emits a constant value through a trainable connection weight, can be present in the input and hidden layers. The purpose of the bias node is to permit the inclusion of a constant value in the trained network model, which is analogous to the constant (y-intercept) of a polynomial function (Bishop, 1995).

Backpropagation (BPNet)

If backpropagation is chosen as the MLP training method, then the system will create and train a BPNet object with the architecture and learning parameters specified by the OGA. The initial connection weights of the BPNet are random numbers between –1.0 and 1.0. As described in *Supervised Learning* above, during a round of BPNet training every member of the training set is presented to the input nodes, and the ANN prediction is compared with the target value. After each case is presented, the connection weights are modified according to the backpropagation algorithm. After each round of training, the members of the test set are presented to the network to allow the measurement of generalization accuracy. This procedure is repeated for a predetermined number of rounds.

Several modifications to the basic backpropagation algorithm have been implemented. These modifications are each represented by parameters that can be optimized by OGA, and in each case there is a value that can inactivate the added feature.

- Momentum adds another term to the function used to carry out connection weight modifications. When momentum is used, the change (Δw) is dependent on the previous node error as well as the current value:

$$\Delta w_{a,b} = -\eta * \delta_{b(n)} * x_a + \mathbf{m} * \delta_{b(n-1)} \qquad (2.5)$$

  Where η is the learning rate, $\delta_{b(n)}$ is the current error of the destination node, $\delta_{b(n-1)}$ is the error calculated for the previous training case, and m is the momentum term that serves as a multiplier for the old error value.
- Learning rate decay affects the learning rate η described above. In early training rounds, a large value of η may be desirable, whereas a lower η value may be desirable in later rounds when the connection weights are closer to optimal values (Hassoun 1995). In GANN, two parameters affecting learning rate decay can be optimized: one determines the training round where learning

rate decay will begin, and the other determines the rate of decrease of η once it begins.

- Weight decay is applied as a counter to overfitting. Where large connection weights often result from overfitting (Krogh and Hertz, 1992), it is desirable to add a 'drive' toward smaller connection weights. The implementation is similar to learning rate decay above: after each training case is presented to the ANN, a small value is subtracted from each connection weight (or added if the weight is negative). As with learning rate decay, the OGA can optimize the training round where rate decay begins, and the proportional decrease in connection weights during training.

Genetic Algorithms (GANet)

A GANet object undergoes training and testing phases as with BPNet above. However, since the GA is similar to reinforcement learning, the connection weights are not modified after each training case is presented. Instead, a cumulative score records whether each training case was predicted correctly or not, assigning a score of 0.0 for each incorrect prediction and 1.0 for each correct prediction. At the end of each training round, this score is divided by the number of training cases to yield a summary value between 0.0 and 1.0, which is recorded as the 'fitness' of the current set of connection weights. Each member of the population will specify different connection weight values, and will be assigned its own fitness value at the end of training. Similar trials on the test set can be carried out to determine generalization ability, but the scores derived from this set do not affect the 'Inner Genetic Algorithm' (IGA) evolutionary operations described below.

Once every parameter set within the population has been assigned a fitness score based on its training set prediction accuracy, the sets are ranked and sorted in descending order according to fitness. Recombination, mutation and migration (as described in *GANN for Parameter Optimization*, above) are then carried out to intermingle the sets of connection weights. While the parameters that affect OGA Chromosome evolution are user-specified, the GANet evolutionary parameters can be optimized by the OGA. An

additional evolutionary feature is implemented for GANet objects: after recombination, a fraction of the inner genetic algorithm (IGA) Chromosomes are compared with each other. The Euclidean distance between vectors of genes is determined for each pair of IGA Chromosomes, and pairs that are closer to one another than a given threshold may be subject to elimination. In such a case, one of the pair of similar Chromosomes is eliminated from the population, and replaced with a new recombinant.

When these evolutionary operations are complete, another round of training and testing can begin. This cycle of evaluation and evolution is conducted for a fixed number of generations that is defined at the beginning of execution.

The maximum number of training rounds (for BPNet) and generations (for GANet) were fixed at 500 and 50, respectively, because these amounts of connection weight optimization were found in trial runs to reach stable classification and prediction accuracy (data not shown). However, training was halted before the maximum number of rounds if no improvement in classification accuracy was seen over a span of 20 training rounds.

Set Definitions

In these experiments, genomic DNA sequences are extracted and assigned to either a 'positive' or 'negative' set based on their position within the genome. While the extraction parameters vary between experiments, the definition of the sets is consistent. Within a single experiment, the extracted sequences in both sets will all be of equal length.

The positive set consists of sequences that are likely to contain transcriptional regulatory features, extracted from upstream of the start codon of ORFs within a genome. An optional constraint is that the upstream sequence be entirely intergenic, and not overlap with coding sequence further upstream. This constraint is important because of the severe compositional difference between intergenic and protein-coding sequence: for instance, in E. coli the average G+C content is ~50% in coding sequence but only ~40% in intergenic sequence (Marìn *et al.*, 1999). A minimum length constraint is imposed to exclude the upstream sequences of downstream operonic sequences from the positive set.

This second constraint is necessary to exclude non-leading operonic genes if the first constraint of non-overlap is not imposed.

The negative set consists of sequences that are unlikely to contain transcriptional regulatory information. These sequences can be derived from intergenic sequence that is not contained within the upstream regions defined above, either further upstream from a start codon than the positive set sequence, or located between two convergently transcribed genes, and therefore downstream from both of them. Protein-coding sequences can also be included in the negative set, but these are less suitable than other intergenic sequences: they will be strongly biased due to selective pressure on amino acid codons, and as mentioned above differ substantially in nucleotide composition from intergenic sequence. While conserved patterns are present throughout the genome, including downstream signals such as transcription terminators (Ray and Daniels, 2003), the composition of protein-coding sequence is more strongly biased due to its conserved function.

Scoring of ANN Predictions

A scoring method is necessary to evaluate and to compare the prediction accuracy of different ANNs. Two scoring methods are used by GANN, though in practice they yield results that are very similar, and only one is typically reported. The scoring methods consider both the neural network prediction for a given set member and the desired prediction determined from the set assignment. Both types of score are calculated by comparing the predicted value(s) derived from the ANN's output node(s) to the target value determined by positive or negative set membership. Since the tanh activation function is used to constrain node values, every output node will have a predicted value and a target value between –1.0 and 1.0.

The *distance score* is obtained by calculating the difference between the predicted and target values, and subtracting the result from 2.0, which is the magnitude of the prediction range of each output node. A distance score of 2.0 therefore corresponds to a prediction that matches exactly the target value, while a score of 0.0 corresponds to a

completely incorrect prediction. Random predictions should yield a mean distance score near 1.0.

The *qualitative accuracy (QA) score* assigns a threshold value for predictions. If the ANN prediction (the output node value) is greater than the threshold, then the prediction is interpreted as positive, while a score less than or equal to the threshold is taken as negative. A score of 1.0 is assigned for each case where the ANN's prediction is correct, while an incorrect prediction is assigned a score of 0.0. The mean QA score for a set of predictions indicates the overall classification accuracy: 1.0 indicates that every sequence was correctly classified, 0.5 indicates random predictions, while 0.0 indicates misclassification of every case. This calculation is consistent with the standard definition of prediction accuracy (Kohavi and Provost, 1998), where the total number of correct predictions (true positives and negatives) is divided by the total number of predictions (true positives and negatives + false positives and negatives). The accuracy scores reported for the GANN experiments in Chapters 3 and 4 are all QA scores.

During the training or testing phase, the ANN will be presented with positive and negative set members. The sizes of the positive and negative sets need not be equal, but better-than-random predictions could be obtained by the ANN in this case by classifying each example as a member of the larger set. For instance, if the positive and negative sets contain 200 and 1800 cases respectively, then an ANN (or any other classification scheme) that predicts every example to be a member of the negative set would achieve 90% classification accuracy. To avoid this effect, the contribution of the positive and negative sets to the overall accuracy score are weighted to ensure that each set contributes 50% of the final training or test score.

Design and Implementation

The core evolutionary component of the GANN algorithm is adapted from software developed in the C++ programming language by Robert L. Charlebois. The Outer Genetic Algorithm (OGA) class is derived from class SelectableBase (Figure 2-2), which contains a number of Evolvable objects, each representing a population of parameter sets. A complete parameter set is implemented as a Chromosome object, with

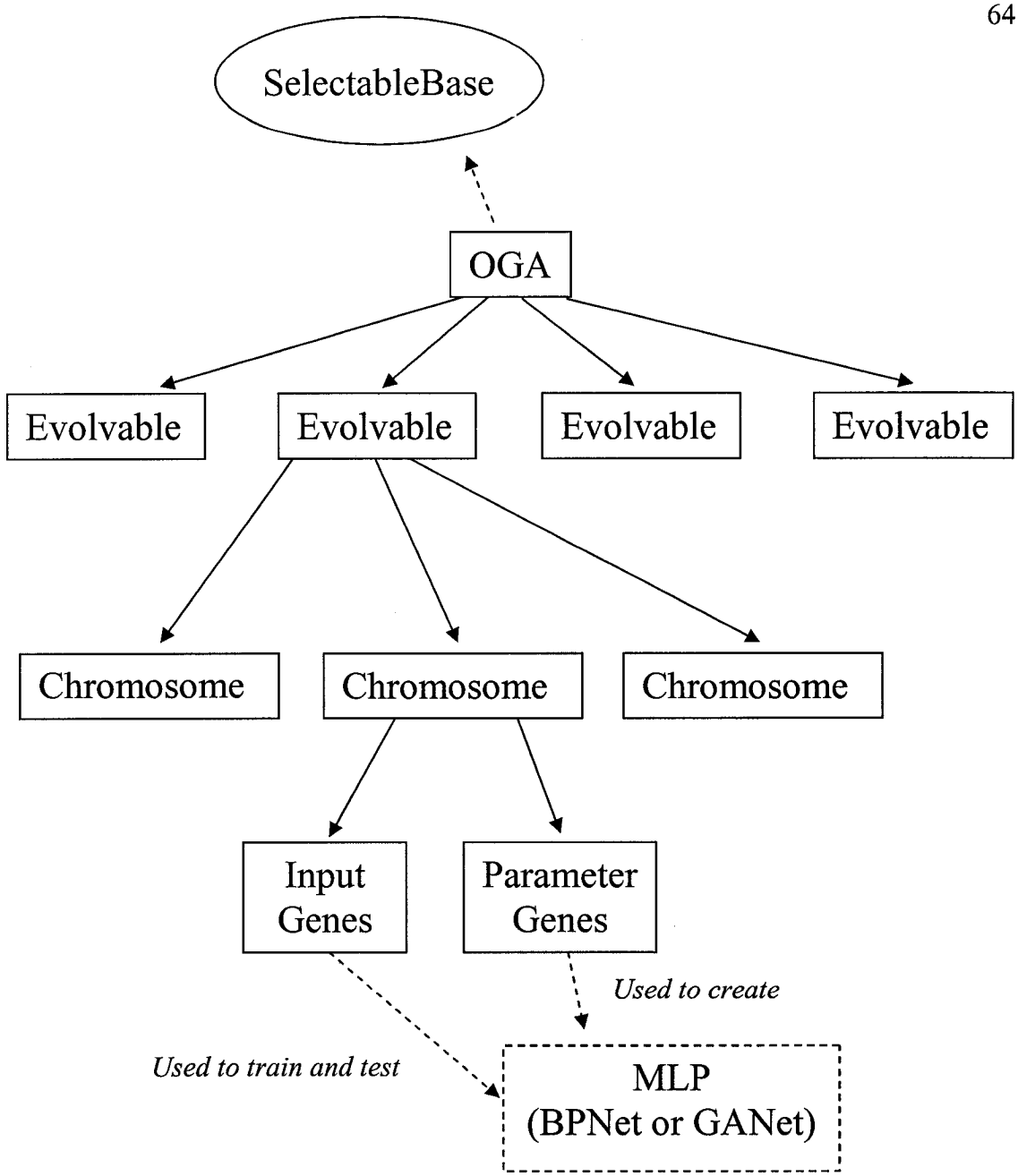


Figure 2-2: Organization of class OGA, which implements the Outer Genetic Algorithm. OGA is derived from class SelectableBase, which contains one or more Evolvable objects, each of which contains in turn multiple OGA Chromosomes. Each OGA Chromosome contains a set of genes that specify the architecture and parameters of a MLP, and another set of genes that specify which indices are used to train the MLP.

each individual parameter represented by a Gene. Two types of Gene are contained within an OGA Chromosome: one type is used to represent neural network parameters, and is represented by a floating-point value; the other identifies which inputs out of the complete set will be used by the neural network, and is arranged as a set of boolean (true or false) values. These boolean values correspond to all of the possible input indices, and the values that are 'true' for a particular Chromosome are the indices that will be considered by neural networks generated from this Chromosome. During the OGA recombination phase, an equal number of 'true' values are extracted from both parents to yield the input set of the offspring.

The set of Genes from an OGA Chromosome are used either to define a single backpropagation neural network, or an IGA-trained neural network. Both types of ANN are derived from the abstract C++ base class MLP (for Multi-Layer Perceptron), and many of the operations (such as forward propagation of input values) are identical for both. Class BPNet implements the backpropagation-specific learning methods, and uses parameters such as learning rate, momentum, weight decay and learning rate decay as described above. Since BPNet objects are initialized with random connection weights, there is an option to perform replicated training runs of the same OGA Chromosome with separate sets of randomly initialized connection weights. Class GANet is derived from classes MLP and SelectableBase (Figure 2-3), thereby combining the neural network architecture with evolutionary operations for training. The organization of GANet is similar to class OGA above, where the connection weights of a single neural network are implemented as Genes of a single IGA Chromosome. A population of interbreeding IGA Chromosomes is organized into an Evolvable, and multiple non-interbreeding Evolvables can be specified within a single SelectableBase object. As with BPNets, the connection weights for a population of GANets are also randomly initialized. However, since each connection weight will have a wide range of starting values within the population, and recombination can create new combinations of these connection weights, a population of GANets is less likely to be influenced by random starting conditions than a single, gradient-trained BPNet. In place of replication, GANet has a parameter that determines the number of IGA Chromosome generalization scores to consider when calculating the fitness of the OGA Chromosome that created them. Thus, the fitness of a given OGA

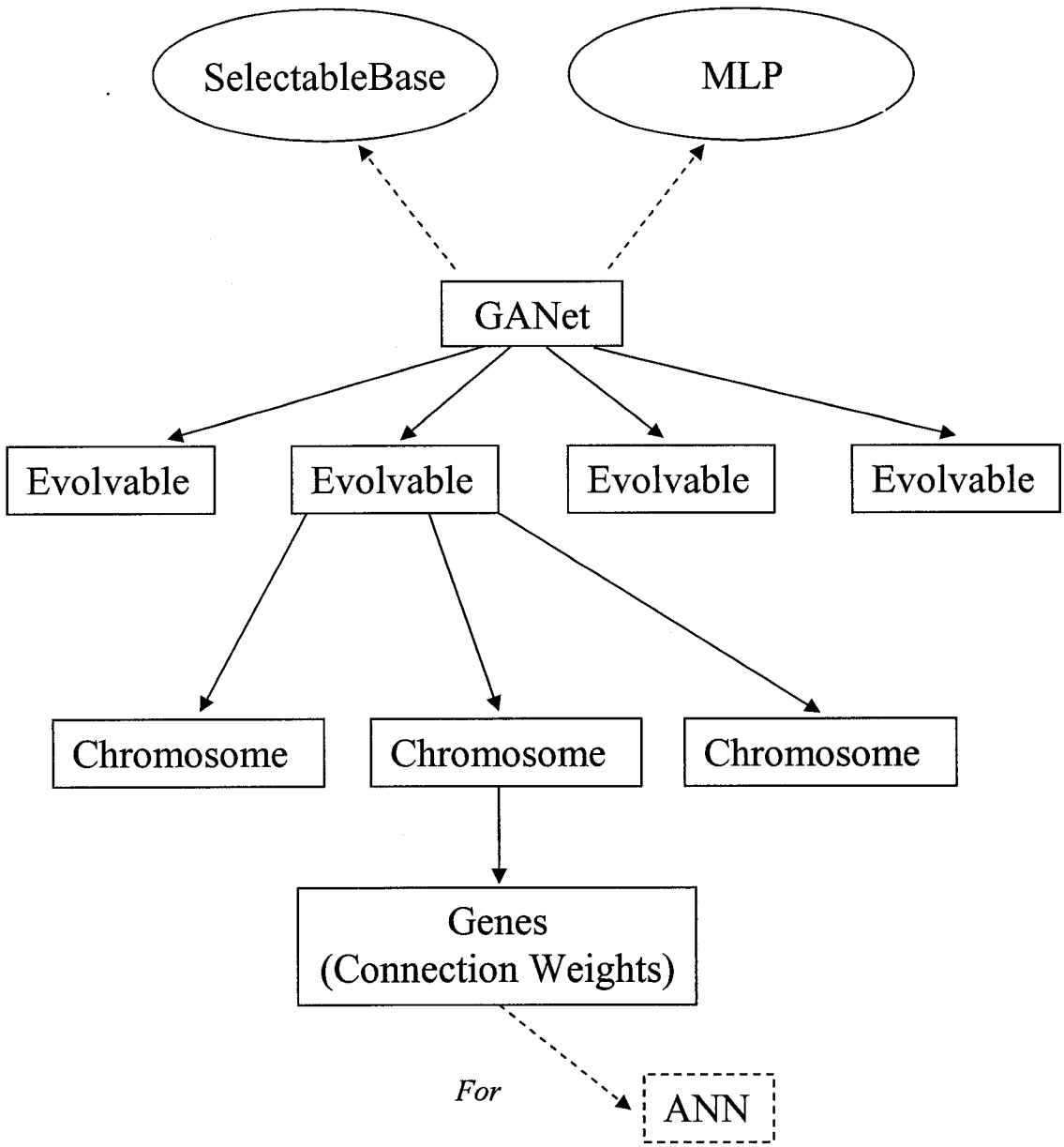


Figure 2-3: Organization of class GANet, which implements the Inner Genetic Algorithm (IGA). Class GANet is derived from classes MLP and SelectableBase. Each GANet object contains one or more Evolvables, each of which contains in turn multiple IGA Chromosomes. The genes contained in a single IGA Chromosome define the connection weights for an artificial neural network.

Chromosome can be derived from only the best few of its ANNs, or from most or all of them.

[3]

Pattern Detection in Regions Upstream of *Escherichia coli* K12 ORFs Using an Interval Representation of DNA

Motivation

Neural network models of DNA have typically used a sequential encoding of bases to represent a stretch of DNA (Wu, 1997). While this encoding has the advantage of preserving context-specific positional information, it usually imposes the requirement for an exact alignment of sequences. An offset of a single nucleotide when a consensus-containing sequence is presented to the neural network will result in consensus positions being represented at the wrong input nodes. If a protein-binding site has a variable-length spacer region between two or more blocks of conserved sequence, then the traditional sequence encoding will not be able to learn more than one spacing class of the site. This limitation was observed in using completely connected MLPs trained with the backpropagation algorithm to detect *E. coli* promoters: networks trained on a specific spacing class performed well, but other spacing classes could not be handled by the same networks. Solutions to this problem include an approach by Mahadevan and Ghosh (1994), who combined the predictions of two separate networks, one trained only on the conserved promoter regions bound by σ, the other trained on a 50 nt promoter region. Promoter2.0, the eukaryotic promoter detection tool developed by Knudsen (1999), trained two-layer perceptron networks to detect conserved features such as the TATA and GC boxes, and allowed the prediction of one perceptron unit to serve as an input to another, thus allowing combinatorial pattern detection without the need for positional representations. Reese (2001) implemented a time-delay neural network (see Chapter 2) architecture to allow variation in spacing between features.

The goal of the present experiment is to use a data encoding scheme that does not depend on properly aligned sequences to represent features in DNA. Instead of encoding

the sequence as a linear string of bases, the encoding used here represents the frequency of pairs of defined bases separated by different lengths of unspecified nucleotide bases. Using GANN as the detection tool, upstream regions extracted from *E. coli* K12 are searched for conserved patterns, represented by these nucleotide interval counts. In a more general sense, these experiments are used to examine the behaviour of GANN on biological data, with a focus on the optimization of ANN parameters by the Outer Genetic Algorithm.

Experimental Design

Sequence Extraction and Set Definition

Unique to this set of GANN runs, the sequence sets used to train the system were dynamically extracted from genomic DNA during training. In addition to the architectural parameters optimized by OGA that are described in Chapter 2, this set of experiments was also designed to optimize the length *L* of the positive and negative set sequences. Upstream regions were defined as the first 150 base pairs (bp) upstream of the start codon of an open reading frame from *E. coli* K12 that did not overlap with any other protein-coding DNA sequences. Negative set sequences were drawn from intergenic regions that did not overlap with defined upstream or protein-coding sequences. In addition to optimizing the length of upstream sequences, GANN also searched for the optimal positioning of the extracted sequence within the 150 bp upstream region. The starting position *P*, which is distal from the start codon, was calculated as follows:

$$P = (150 - L) * S \qquad (3.1)$$

Where S is a real-numbered value between 0.0 and 1.0, representing where in the range of possible start points to extract the sequence. Any fractional component of P was rounded down to yield an integer value.

For each run of GANN, 200 positive sequences and 200 negative sequences were extracted at random from the *E. coli* K12 genome. Duplication of sequences within the

set was avoided by checking the physical distances between members of each set: a new sequence could only be included in the positive or negative set if it was more than 150 bp distant from every other sequence already in that set. Once assembled, these sets were split in half, with 100 members from each set used for training and 100 for testing.

Sequence Encoding

This experiment relied on nucleotide sequence alone, to permit testing of the system with a reasonably small number of indices. The interval representation yields 16 different counts for each length of spacer that is considered, for instance, an interval size of 1 will yield counts for the trinucleotides ANA, ANC, ANG, ANT, CNA, CNC, CNG, CNT, GNA, GNG, GNT, TNA, TNC, TNG, and TNT. Including structural definitions such as helical twist and deformation energy would yield thousands of indices, which may be too many to deal with in a single run of GANN. Also, since some GANN runs in this experiment considered the entire pool of indices at once rather than subsamples from the pool (see *GANN Runs and Controls* below), including structural indices would have yielded giant networks that could not be trained in a reasonable amount of time.

The maximum size of sequence interval to consider was optimized by OGA. Intervals of size 0, 1, and 2 were always included, but more sizes up to a maximum of 14 could be added, depending on the value of the relevant variable in the current OGA Chromosome. Since each interval size yielded 16 feature counts, the total number of indices, and therefore the number of ANN input neurons, was equal to 16 times the number of interval sizes considered. The number of input nodes thus ranged between 48 and 240 in these experiments. In situations where index pool subsampling was performed, each OGA Chromosome was randomly assigned a set of indices from the pool to use as input. When each OGA Chromosome was first created, there was a probability of 0.25 that a specific index would be included; this resulted in Chromosomes that specified ¼ of the available (48 to 240) indices on average.

GANN Runs and Controls

Several different experimental runs and controls were performed, to study the performance of different components of GANN. The 18 separate runs that were performed are summarized in Table 3-1. Within the experimental runs, the key distinctions were in the use of BP or GA as the training algorithm and the use or absence of subsampling from the index pool performed by OGA. Experimental and control runs were performed with three types of sequence:

Real (Experimental): DNA sequence extracted from a file containing the complete genome of *E. coli* K12, and a file containing a table of annotated ORFs. The table of ORFs was extracted by performing a 'Sorted Lists of Protein Characteristics' query for *E. coli* K12 at the NeuroGadgets Inc. Bioinformatics Web Service (www.neurogadgets.com), with the ORF start point within the chromosome of *E. coli* K12 as the sorting criterion.

Shuffled, non-spiked (Negative S): DNA sequence derived from the *E. coli* K12 genome, but with randomized order of nucleotide bases. The exact nucleotide composition was preserved, but the order of bases was shuffled to disrupt motifs and local sequence biases.

Shuffled, spiked (Positive S): Shuffled as above, but after shuffling, sequences from a set of characterized promoters obtained from O.N. Ozoline (Institute of Cell Biophysics, Russian Academy of Sciences) were inserted into the candidate upstream regions. Each promoter from this set has a characterized transcription start site, which was inserted at the end of the 150-bp sequence which corresponds to the translation start site. Since the distance between conserved promoter motifs and the transcription start site is consistent, this positioning will align the known promoter sequences in the positive data set.

Table 3-1: Summary of GANN runs performed to detect patterns in sequence interval data. The ID is a shorthand form that is used to identify the different runs in this chapter. The learning algorithm used was either backpropagation (BP) or an 'inner' genetic algorithm (GA). The OGA Chromosome specified the indices used by the ANN in one of two ways: through random samples of indices taken from the index pool (Input Vector Optimization), or by specifying the number of nucleotide intervals to consider. The different types of sequence used and the run categories are described in the **Experimental Design** section of this chapter.

ID	Category of run	Type of Sequence	Learning Algorithm	Input Vector Optimization?
Er1	Experimental	Real	GA	No
Er2	Experimental	Real	BP	No
Er3	Experimental	Real	GA	Yes
Er4	Experimental	Real	BP	Yes
Ps1	Positive S	Shuffled, spiked	GA	No
Ps2	Positive S	Shuffled, spiked	BP	No
Ps3	Positive S	Shuffled, spiked	GA	Yes
Ps4	Positive S	Shuffled, spiked	BP	Yes
Pr1	Positive A	Real	GA	No
Pr2	Positive A	Real	BP	No
Pr3	Positive B	Real	GA	No
Pr4	Positive B	Real	BP	No
Ns1	Negative S	Shuffled, non-spiked	GA	No
Ns2	Negative S	Shuffled, non-spiked	BP	No
Nr1	Negative A	Real	GA	No
Nr2	Negative A	Real	BP	No
Nr3	Negative B	Real	GA	No
Nr4	Negative B	Real	BP	No

In addition to the controls described above, two more types of positive and negative controls were derived from the **Real** sequence data described above. The positive controls tested the sensitivity of the system to signals that should easily be detected, while the negative controls were used to determine the ability to learn and to predict signals in data that should contain none.

Positive Control 'A': A trivially easy positive control was performed to test the learning ability of GANN. In this control, every member of the positive sequence set was replaced with a single stretch of adenine nucleotides, while every member of the negative sequence set was replaced with a single stretch of guanines. These two patterns should stimulate only neurons associated with the patterns $A(N_k)A$ and $G(N_k)G$ respectively, and should be easily classified.

Positive Control 'B': To test the effect of overrepresenting a strong pattern in real sequence, another control was performed as in 'A' above, but with only 20% of the total sequence in each positive and negative set member replaced by a string of consecutive adenines or guanines. This string of nucleotides was inserted at a random position in each sequence, with permitted start points anywhere from position 1 to position 131. While the signal associated with the inserted patterns now has to compete with the noisy signal from the surrounding sequence, a trained ANN should still be able to differentiate between positive and negative set members without much difficulty.

Negative Control 'A': The training set, containing 100 positive and 100 negative cases, was shuffled to yield new 'positive' and 'negative' sets that each contained 50 upstream and 50 non-upstream sequences. In this instance the *set membership* rather than the *sequence* of each member that was shuffled to yield the new sets. This control was used to demonstrate the ANN's ability to discriminate between sets that should have no characteristic patterns that separate them.

Negative Control 'B': In a separate set of runs, the positive and negative members of the *test* set were shuffled as in negative control 'A' above. The ANN would be able to learn any patterns present in the training set, but any patterns associated exclusively with the positive training set would then be found in both the positive and negative test sets. Thus, even if the ANN performed well on the training set, poor prediction accuracy would still be expected from the test set.

The set of parameters optimized by the OGA are shown with their permitted ranges in Table 3-2. Boolean values (true and false) were represented as real-numbered values between 0.00 and 1.99, with any value < 1.00 interpreted as 'false', and any value ≥ 1.00 interpreted as 'true'. Depending on whether BP or GA was used as the training algorithm, only the set of variables identified with 'BP Learning' or 'GA Learning' were used. The values of the variables InputScale, NumEvo, MigRate, and Conj had no impact on the present GANN runs, and could therefore demonstrate the population dynamics of variables under no selection.

Thirty OGA training rounds were carried out for each run, with a population size of 50 OGA Chromosomes when subsampling of the index pool was performed, and 25 otherwise. BP and GA nets were both trained for 500 'inner' rounds, with the training and test set scores recorded after the final round.

Results

Prediction Accuracy

'Easy' positive controls were performed to verify the ability of GANN to perform the task of classification without the need to address difficulties presented by real biological motifs. Figure 3-1 shows the results of the four positive control runs with real sequence, where long homogeneous repeats were introduced into the training and test sets. In the first two runs (Pr1 and Pr2), where the entire sequence was replaced with a tract of mononucleotides, the mean training and test scores were close to 1.0 in the first few OGA generations, and both the GA- and BP-trained populations quickly

Table 3-2: Summary of parameters optimized by the OGA, with maximum and minimum values. Among parameter types, the 'Sequence', 'Data Representation' and 'Architecture' types are described in the Methods section of the current chapter, while the 'GA Learning' and 'BP Learning' types are described in Chapter 2. Abbreviations for each parameter type used in this chapter are shown, along with the minimum and maximum parameter values permitted by GANN. An asterisk indicates a feature that was not implemented or had no effect on the ANN runs. Each type of input index was standardized over all members of the positive and training set, because the value of the InputScale variable could never decrease below 1.00. The number of IGA Evolvables was always constrained to be between 1.0 and 1.5, which always yielded 1 when rounded down. Since migration occurs between Evolvables, the absence of multiple Evolvables made the MigRate parameter redundant. Finally, the conjugate-gradient learning variant of backpropagation was never implemented, but the parameter controlling its implementation was allowed to persist as another 'unimportant' variable.

Parameter Type	Parameter	Abbreviation	Minimum Value	Maximum Value
Sequence	Window Size	WinSize	50	150
Sequence	Sequence Start Position	StartLoc	0.00	1.00
Data Representation	Maximum Number of Intervals	IntervalNum	3	15
Data Representation	*Standardize input vectors	InputScale	1.00 (true)	1.99 (true)
Architecture	Number of Hidden Nodes	HidNodes	2	50
Architecture	Cutoff between positive and negative predictions	Cutoff	0.30	0.70
GA Learning	Fraction of population subject to mutation	MutProp	0.01	0.90
GA Learning	Rate of mutation within chromosomes	MutRate	0.01	0.99
GA Learning	Maximum magnitude of mutation	MutAmt	1.01	1.80
GA Learning	Proportion of parents that recombine	RecRate	0.10	0.90
GA Learning	Proportion of population replaced by recombinants	RecRepl	0.30	0.95
GA Learning	Proportion of population subject to similarity-based elimination	KillFrac	0.10	0.85
GA Learning	Maximum Euclidean distance between 'similar' Chromosomes	KillDiff	0.10	0.9
GA Learning	Number of GANet Chromosomes	NumChr	100	500
GA Learning	*Number of Evolvables	NumEvo	1.00	1.50
GA Learning	*Migration Rate	MigRate	0.10	0.90
BP Learning	Learning Rate	LearnRate	0.01	1.00
BP Learning	Momentum	Momentum	0.00	0.9
BP Learning	Weight Decay	WtDec	0.00	0.99
BP Learning	Weight Decay start	WtDecStart	0.00	1.00
BP Learning	Learning Rate Decay	LRDecay	0.00	0.50
BP Learning	Learning Rate Decay start	LRDecayStart	0.00	1.00
BP Learning	*Conjugate-Gradient Learning	Conj	0.00	1.00

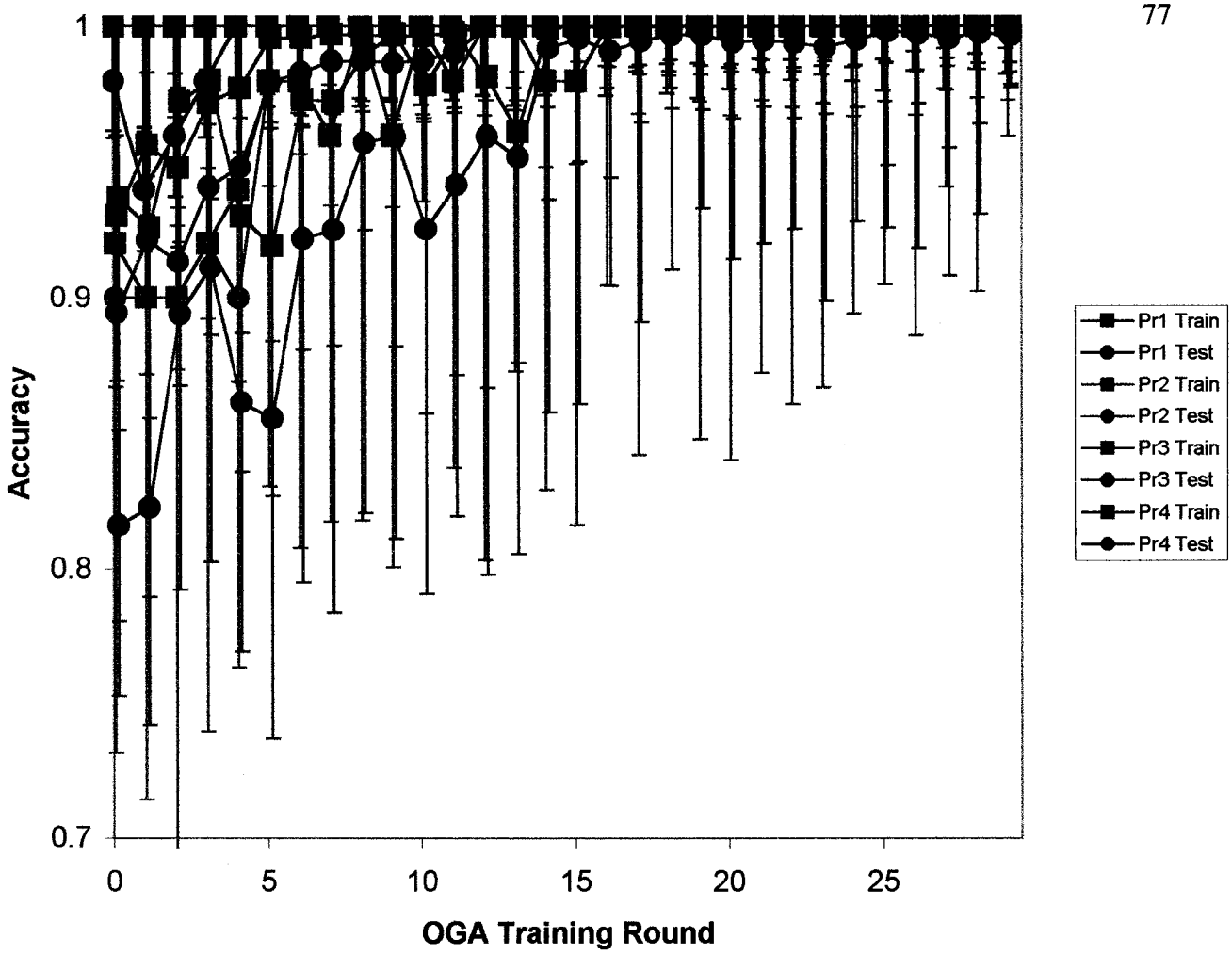


Figure 3-1: Progress in training (filled squares) and test (filled circles) accuracy over 30 OGA rounds for 'Pr' runs (positive controls with real sequence). Different colours of lines, symbols and standard deviation bars indicate different runs, which are described in Table 3-1. In these runs, upstream and non-upstream sequences from *E. coli* K12 were 'spiked' with continuous mononucleotide runs, with either 100% (Pr1 and Pr2) or 20% (Pr3 and Pr4) of the 150 nt upstream sequence replaced. Neural networks in the odd-numbered runs (Pr1 and Pr3) were trained with a genetic algorithm, while those in the even-numbered runs (Pr2 and Pr4) were trained using backpropagation. The scores shown here are the averages over all OGA Chromosomes ± the standard deviation in the current round of training. The scores contributed by an individual OGA Chromosome are either the training and test set scores in the final round of backpropagation training, or the best scores from the population of inner genetic algorithm networks after the final round of crossover.

homogenized to yield an entire population with a prediction accuracy of 1.0. Poor ANN parameter combinations that could not learn the simple pattern were quickly eliminated from the population. In the second pair of runs (Pr3 and Pr4), where only 20% of the sequence was replaced with a repeated mononucleotide, the mean OGA Chromosome score reached 1.0 within the 30 OGA training rounds. However, more OGA generations were required to achieve the maximum mean generalization score, with 12 generations required for Pr3 and 25 for Pr4, versus 4 for Pr1 and 8 for Pr2. While training scores reached the maximum more quickly in the BP runs, GA required fewer OGA rounds to reach the best generalization scores.

The experimental runs, summarized in Figure 3-2, illustrate a dramatic difference between supervised learning and evolutionary methods. The BP runs, either with (Er4) or without (Er2) input vector optimization, were able to classify the training set with a high degree of precision, reaching accuracy scores of 1.0 in later rounds. However, the patterns learned from the training set were clearly not representative of the test examples, because mean generalization over all OGA Chromosomes never reached even a score of 0.6, with the best OGA Chromosomes yielding scores between 0.63 and 0.65 (Figure 3-6, later in this chapter). The GA runs achieved similar mean generalization scores of slightly less than 0.6, with the best OGA Chromosomes again in the 0.63 – 0.65 range. However, the GA trained networks were not able to learn the training set perfectly, with mean accuracy scores around 0.8. The training scores also decreased slightly over the 30 rounds of OGA training, in sharp contrast with the increases seen in the BP populations. The use of input vector optimization did not yield increased prediction accuracy, but did produce smaller networks with a similar range of prediction accuracy, thus reducing the required number of inputs and speeding up execution time.

The discrepancy between training scores associated with BP and GA networks was also observed when randomized sequence was used to generate the data sets (Figure 3-3). Even with randomized data, the BP-trained networks (Ns2) were eventually able to achieve near-perfect prediction accuracy on the training set. Since there should be no conserved patterns within either the positive or negative training sets, this outcome strongly suggests that the ANNs are memorizing each training case individually. The large size of the neural network relative to the size of the training set (see *Parameters*

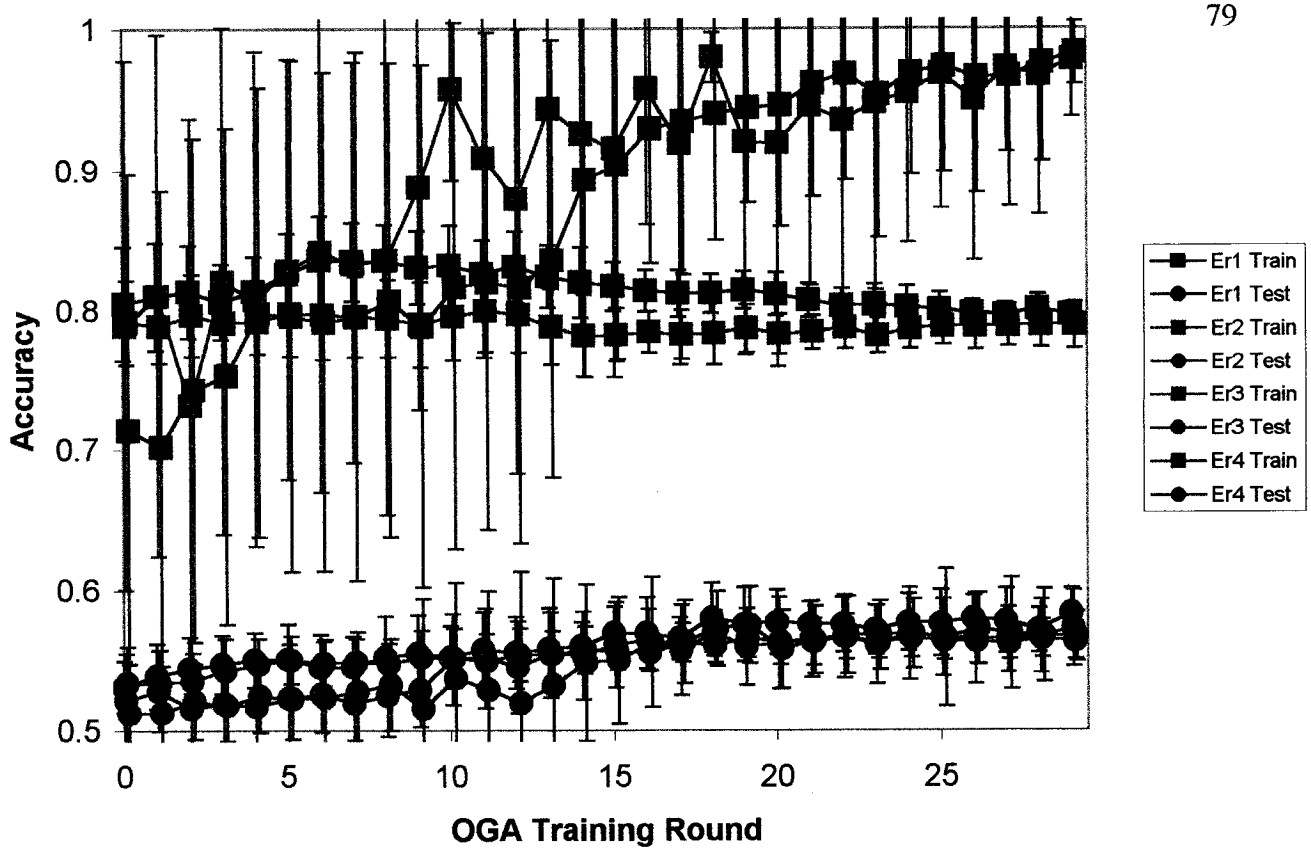


Figure 3-2: Progress in training (filled squares) and test (filled circles) accuracy over 30 OGA rounds for 'Er' runs (experimental runs with real sequence). Different colours of lines, symbols and standard deviation bars indicate different runs, which are described in Table 3-1. In these runs, upstream and non-upstream sequences from *E. coli* K12 were used to train and test ANNs, either with (Er3 and Er4) or without (Er1 and Er2) input vector optimization. Neural networks in the odd-numbered runs (Er1 and Er3) were trained with a genetic algorithm, while those in the even-numbered runs (Er2 and Er4) were trained using backpropagation.The scores shown here are the averages over all OGA Chromosomes ± the standard deviation in the current round of training. The scores contributed by an individual OGA Chromosome are either the training and test set scores in the final round of backpropagation training, or the best scores from the population of inner genetic algorithm networks after the final round of crossover.

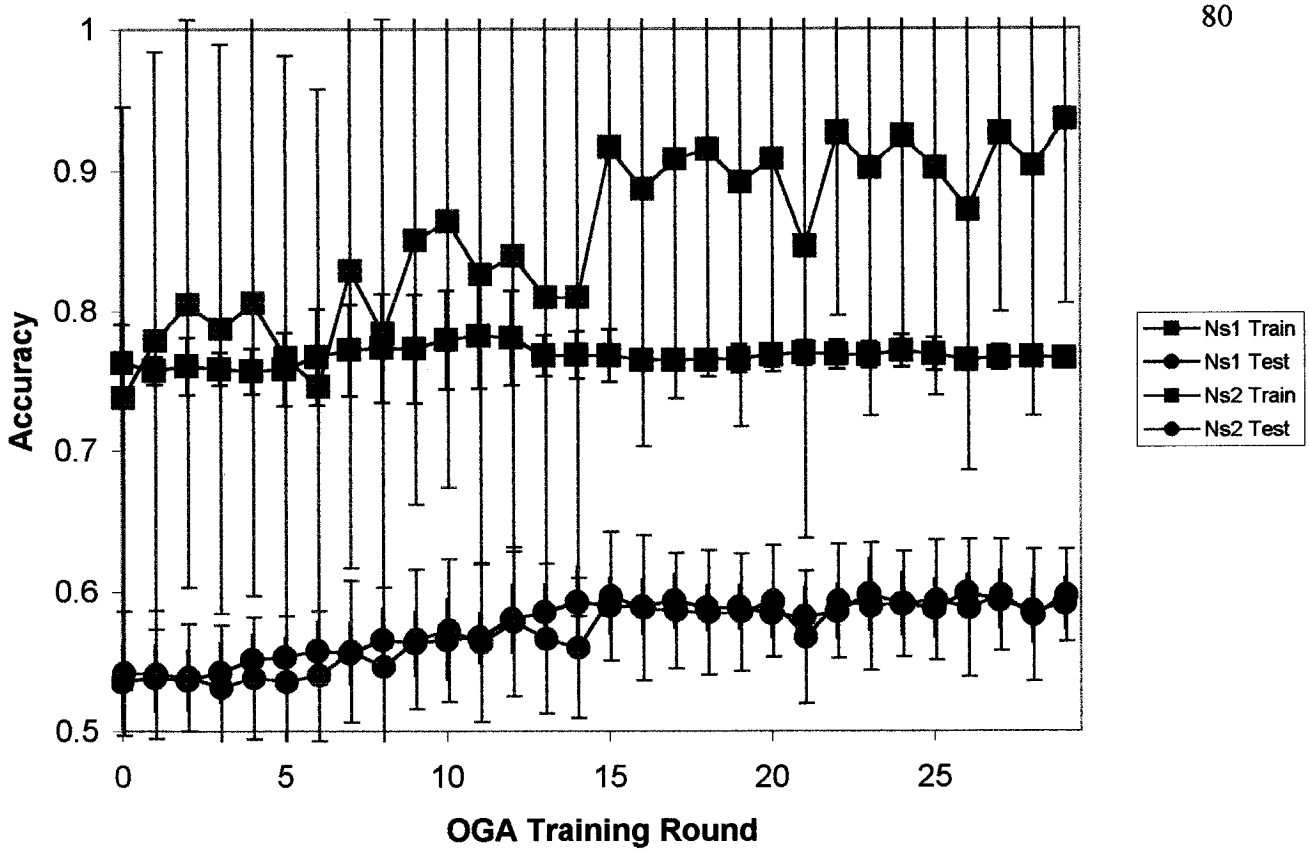


Figure 3-3: Progress in training (filled squares) and test (filled circles) accuracy over 30 OGA rounds for 'Ns' runs (negative control runs with shuffled sequence). Different colours of lines, symbols and standard deviation bars indicate different runs, which are described in Table 3-1. In these runs, upstream and non-upstream sequences were extracted from a copy of the *E. coli* K12 with randomized nucleotide order. Neural networks in the odd-numbered run (Ns1) were trained with a genetic algorithm, while those in the even-numbered run (Ns2) were trained using backpropagation. The scores shown here are the averages over all OGA Chromosomes ± the standard deviation in the current round of training. The scores contributed by an individual OGA Chromosome are either the training and test set scores in the final round of backpropagation training, or the best scores from the population of inner genetic algorithm networks after the final round of crossover.

Optimized by GANN, below) is the most likely cause of the high training scores. The GA-trained networks (Ns1) again yielded training set accuracy of about 0.8. The generalization scores associated with both of these negative controls were slightly higher than the test set scores of the experimental runs. Since the experimental and negative control scores are effectively equal, there is no evidence that any biological patterns were discovered in the experimental runs.

The positive control runs performed on shuffled sequence 'spiked' with promoters (Figure 3-4) yielded training and test set accuracy that is similar to the scores from the experimental runs. The two BP runs (Ps2 and Ps4) were able to learn the training set with an accuracy greater than 0.9, but this accuracy on the training set did not lead to good generalization on the test set, where scores again were around 0.6. The GA runs (Ps1 and Ps3) produced a training set accuracy of about 0.8, with test scores in the same range as seen with the BP-trained networks. Since the accuracy seen here is not different from the accuracy obtained with the negative controls, it is unlikely that the true promoter signals were learned. The scores obtained when input vector optimization was performed (Ps3 and Ps4) were slightly higher than the two other runs. For instance, in the final round of OGA training, the generalization scores obtained with IVO were 0.02 to 0.03 greater than those obtained without IVO.

The negative controls that randomized the membership of either the training set (Nr1 and Nr2) or the test set (Nr3 and Nr4) are presented in Figure 3-5. In contrast with other runs involving backpropagation, the BP algorithm was not able to learn the training set with a high level of accuracy, instead peaking with an average OGA Chromosome score of 0.73. However, this run yielded a generalization score of 0.59 in the final round, which was the highest of the four runs shown in this figure. BP-trained nets were able to learn a non-shuffled training set with accuracy approaching 1.0, but generalization scores on the randomized training set were very low, peaking below 0.52. The GA nets performed equally well on the real and randomized training sets, and the generalization scores were both between 0.55 and 0.56 in the final round.

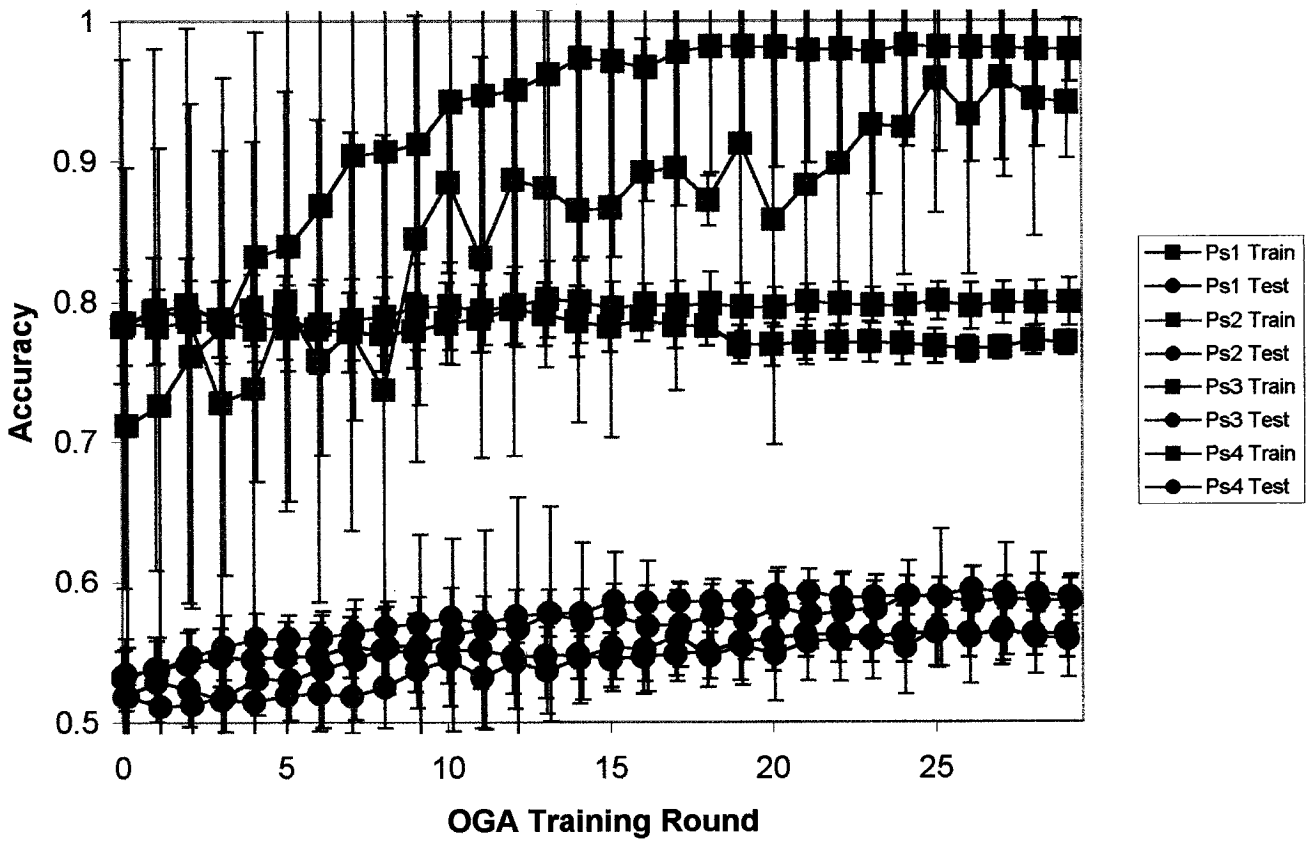


Figure 3-4: Progress in training (filled squares) and test (filled circles) accuracy over 30 OGA rounds for 'Ps' runs (positive control runs with shuffled sequence). Different colours of lines, symbols and standard deviation bars indicate different runs, which are described in Table 3-1. In these runs, upstream and non-upstream sequences were extracted from a copy of the *E. coli* K12 with randomized nucleotide order as in the 'Ns' runs above, but experimentally characterized promoters were then inserted into the positive training and test sequence sets. Neural networks in the odd-numbered runs (Ps1 and Ps3) were trained with a genetic algorithm, while those in the even-numbered runs (Ps2 and Ps4) were trained using backpropagation. The scores shown here are the averages over all OGA Chromosomes ± the standard deviation in the current round of training. The scores contributed by an individual OGA Chromosome are either the training and test set scores in the final round of backpropagation training, or the best scores from the population of inner genetic algorithm networks after the final round of crossover.

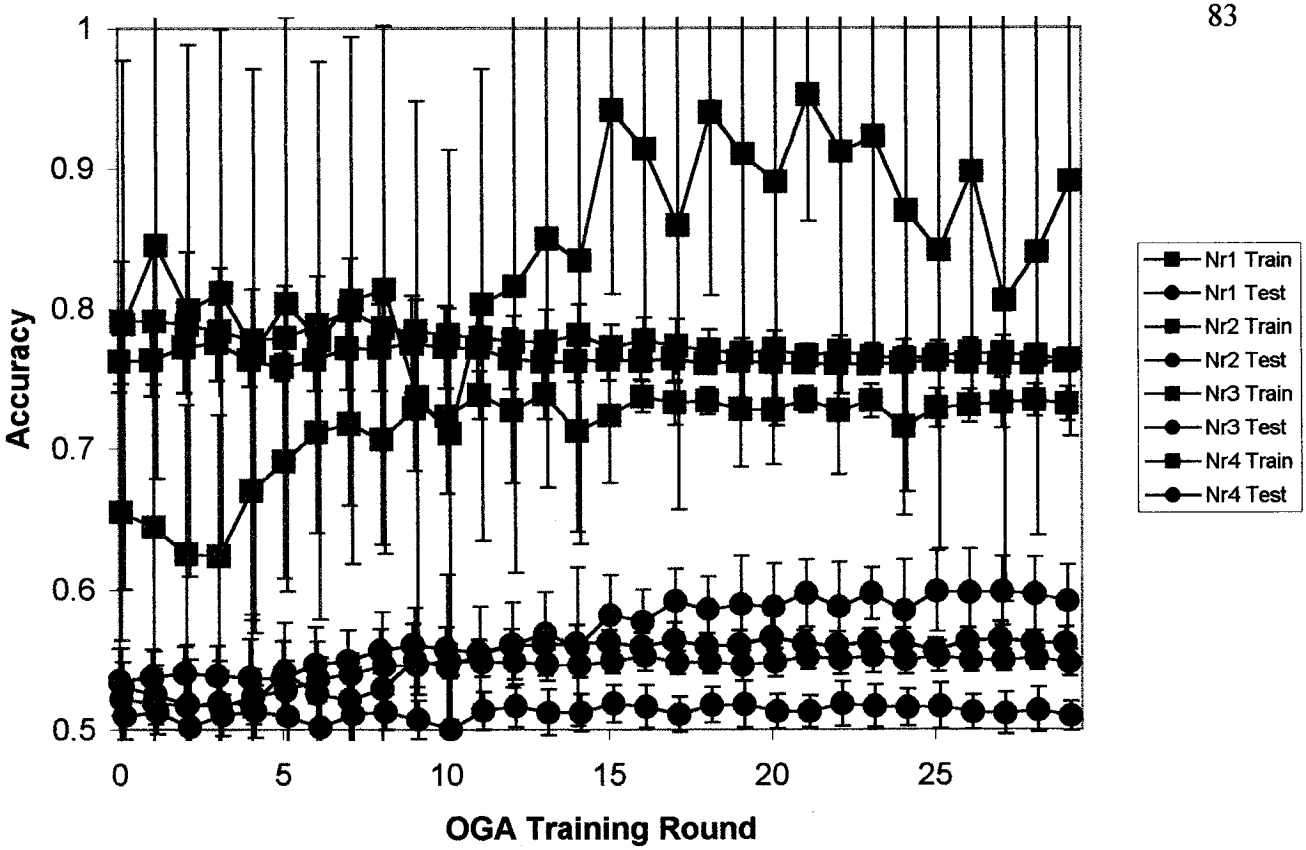


Figure 3-5: Progress in training (filled squares) and test (filled circles) accuracy over 30 OGA rounds for 'Nr' runs (negative control runs with real sequence). Different colours of lines, symbols and standard deviation bars indicate different runs, which are described in Table 3-1. In these runs, the members of either the training set (Nr1 and Nr2) or the test set (Nr3 and Nr4) were randomly assigned to either the positive or negative set. Neural networks in the odd-numbered runs (Nr1 and Nr3) were trained with a genetic algorithm, while those in the even-numbered runs (Nr2 and Nr4) were trained using backpropagation. The scores shown here are the averages over all OGA Chromosomes ± the standard deviation in the current round of training. The scores contributed by an individual OGA Chromosome are either the training and test set scores in the final round of backpropagation training, or the best scores from the population of inner genetic algorithm networks after the final round of crossover.

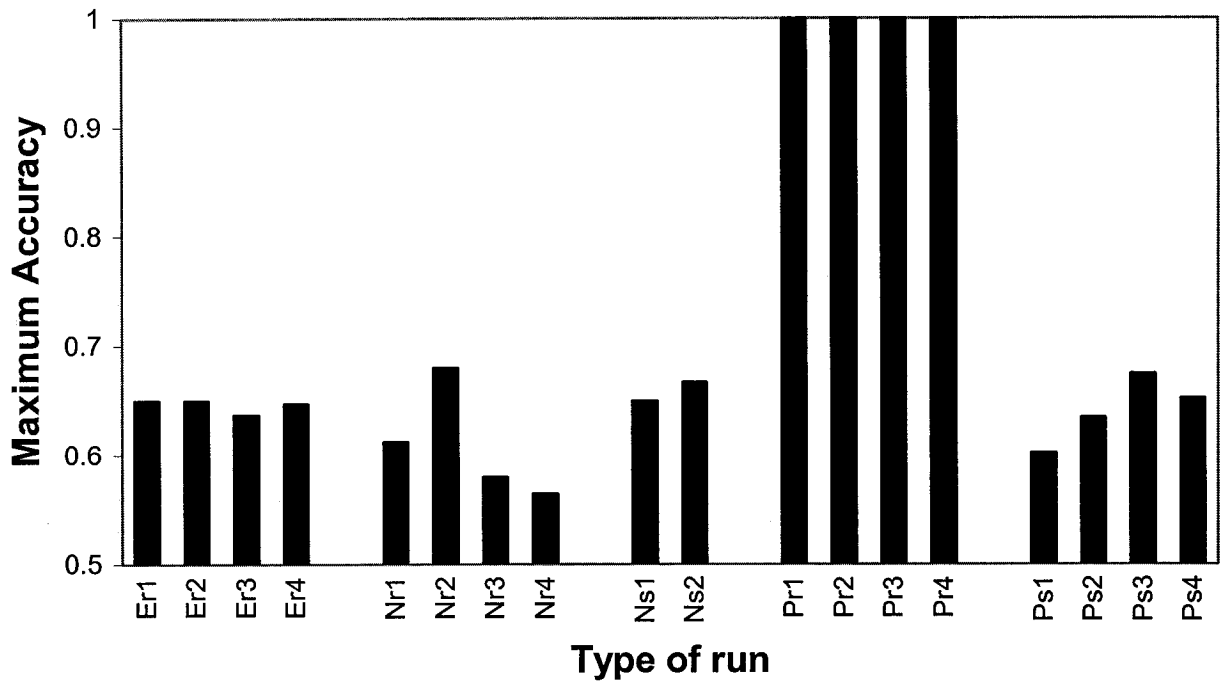


Figure 3-6: Generalization score of the best OGA Chromosome recorded during 30 rounds of training. The scores shown here represent the best accuracy score obtained from a single OGA Chromosome in any of the 30 OGA generations. The score considered for individual OGA Chromosome are either the test set score in the final round of backpropagation training, or the best score from the population of inner genetic algorithm networks after the final round of crossover. The GANN runs are identified by type (E = experimental, N = negative control, P = positive control), sequence used (r = real, s = shuffled), learning algorithm (1,3 = genetic algorithm; 2, 4 = backpropagation) and whether input vector optimization was used (1, 2 = no; 3, 4 = yes).

Parameters Optimized by GANN

The parameters optimized by GANN, shown in Table 3-2, were selected based on their contribution to ANNs that yielded good generalization scores. Analysis of successful parameter combinations and the change in the population structure over time can show which parameters and parameter values were favoured. However, selective pressure is not the only factor that can lead to change in a mean parameter value: since selection acts on whole sets of parameters, some of these may be 'selected' because of their association with good values of important parameters. Especially in a small population, this 'hitchhiking' effect can yield a false impression of favouring a small range of values for an unimportant parameter. There is no explicit linkage in this genetic algorithm, so the contribution of genes from each parent is randomly determined at the time of recombination. The absence of linkage reduces but does not eliminate the possibility of hitchhiking. Since mutation is applied to OGA Chromosomes, genetic drift of unimportant parameters can also lead to changes in the mean value of an unimportant parameter. A strong shift in mean parameter value coupled with a reduction in the range are suggestive of selection, with further support if the behaviour of a parameter was consistent between different GANN.

To assess the significance of parameter change over 30 OGA generations, parametric *t*-tests were used to compare the initial mean value of each parameter with the final mean after 30 rounds of OGA training. The *t*-test describes the statistical significance of the difference in means with a *p*-value (Zar, 1999). Since the sample size in these experiments is only 25 or 50, the reported *p*-value may be sensitive to non-normality of the data points, or unequal variance in the two sets being compared. However, this sensitivity is unlikely to decrease the *p*-value by orders of magnitude, so the relative magnitude of changes in different parameter values can still be assessed by comparing logarithmic transformations of the *p*-values.

Tests were performed to assess the statistical significance of the change in parameters optimized by OGA over 30 training rounds. The results are derived from *t*-tests that compare the distribution of values for a given parameter in the first and final rounds of OGA training. The data presented below reflect the order of magnitude of each

resulting p-value, expressed through a logarithmic transformation. Different shades of red indicate increasing parameter values, different shades of blue indicate decreasing parameter values, and white boxes indicate a minimal change in either direction, with $p > 0.1$. For the purposes of interpretation below, this result will be interpreted as 'no change'. Since the p-values obtained here are not judged against a critical α-value, no correction for multiple tests such as Bonferroni (Zar, 1999) was performed.

The significance of changes in parameters common to all runs are presented in Figure 3-7. Since the range of values of the InputScale parameter were constrained between 1.00 and 1.99, input scaling was always performed, and changes to this parameter had no impact on the functioning of the ANN. In 8 out of 18 runs, the p-value of the parameter change was > 0.10, and 15 out of 18 runs had an associated p-value that was greater than 0.001. The strong changes in the other three runs must be due to either drift or hitchhiking. There was also no clear preference for an increase or decrease in the value of this parameter over time, with 4/10 decreasing and 6/10 increasing over the 30 training rounds. The cutoff value, used to determine whether a prediction is positive or negative, showed a similar distribution of changes, suggesting that this parameter is unimportant to the overall performance of the ANN. No change was observed for 9/18 runs, while 6 showed a decrease and 3 showed an increase. As with the InputScale parameter, a few strongly significant changes were observed, showing that a large directional change does not necessarily imply strong selection for the resulting values.

Within the six parameters shown in Figure 3-7, the one showing the strongest evidence of directional selection is the HidNodes parameter. The value of this parameter increased in 17 out of 18 runs, and the associated p-value was less than 10^{-6} in 11 of those runs. The WinSize parameter showed a strong bias as well: 12 out of 18 showed a decrease, with an associated p-value < 10^{-6} in 9 of those cases. Within the Er and Ps groups, where biological signals from *E. coli* were sought, 8 out of 8 runs showed a decrease in window size.

Weaker trends were observed in the two remaining parameters. The StartLoc parameter increased in 10/18 runs, with an associated p-value < 10^{-6} in half of those changes. However, three runs showed a *decrease* of the same magnitude, and there was little consistency within different groups of runs. The IntervalNum parameter, which

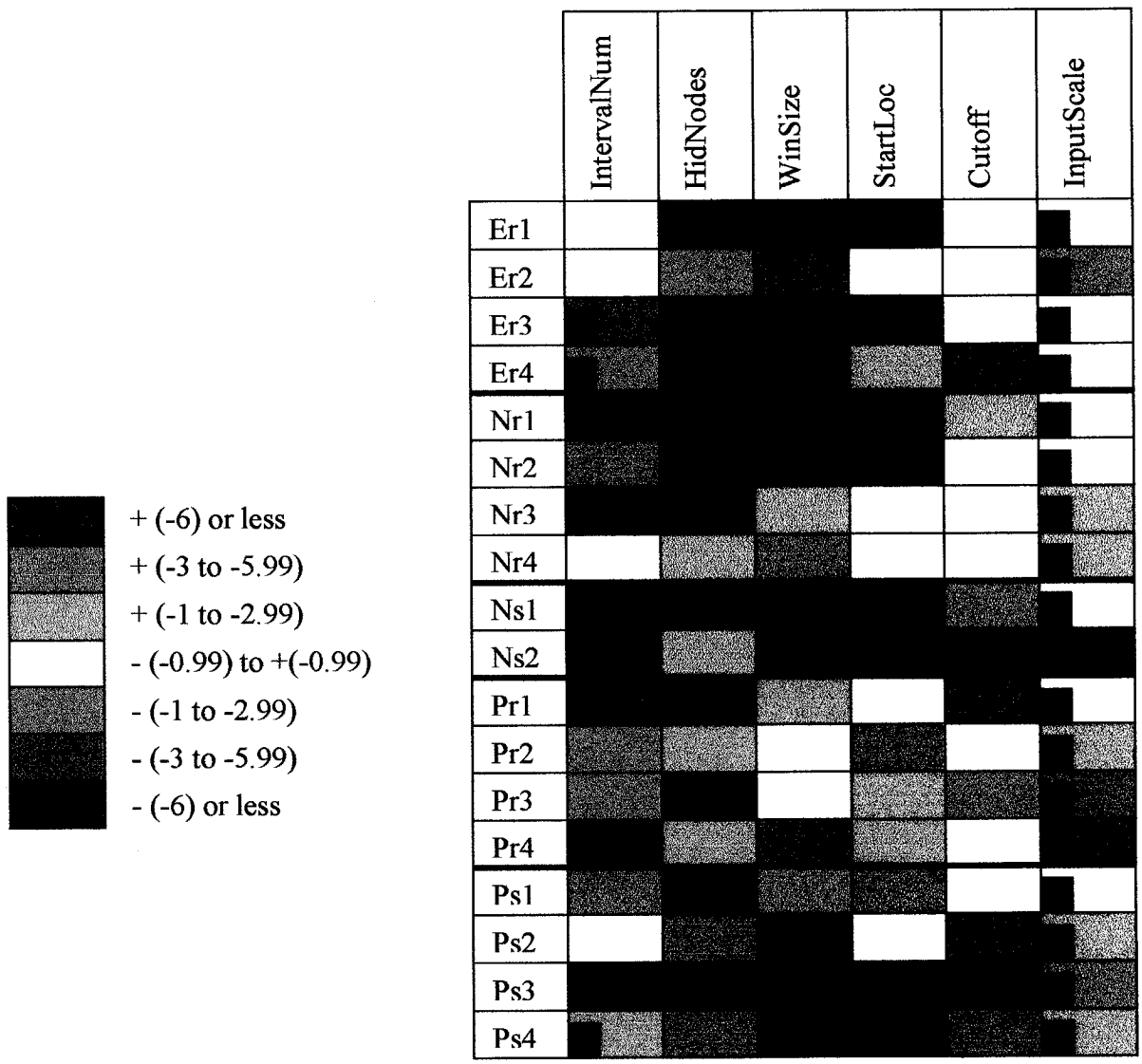


Figure 3-7: Statistical significance of change in parameters optimized by OGA (parameters shared between GA and BP). The colour of each box is determined by the base 10 logarithm of the *p*-value associated with the change in the parameter value over 30 OGA generations. The darkest shade of red indicates an increase with a *p*-value of less than 10^{-10}, while the darkest shade of blue indicates a decrease of the same magnitude. Inset black boxes indicate parameters known to have no impact on ANN performance. The GANN runs are identified by type (E = experimental, N = negative control, P = positive control), sequence used (r = real, s = shuffled), learning algorithm (1,3 = genetic algorithm; 2, 4 = backpropagation) and whether input vector optimization was used (1, 2 = no; 3, 4 = yes).

determined the number of inputs to the neural network in the absence of input vector optimization, showed no consistent increasing or decreasing trend, even though many observed changes had lower associated *p*-values than were observed for both InputScale and Cutoff. In particular, the four cases where IntervalNum did not determine the number of inputs all showed some degree of increase or decrease. Since there is no clear preference for one direction over another and there is no preferred pattern within subcategories of runs, the role of IntervalNum is unclear.

The patterns of change in BP-specific parameters are shown in Figure 3-8. A single parameter (Conj) did not affect the execution of the ANN, and no change was observed in 6 out of 9 cases. The WtDecStart parameter, which defined when in the training cycle weight decay should start, showed a similar lack of change, with 7/9 runs exhibiting no change at all.

Three parameters, all affecting the learning rate, showed biased patterns of change in most runs. The learning rate decreased in 7 of 9 runs, with an associated p-value $< 10^{-3}$ in all four members of the Er and Ps groups. The increase in the learning rate decay term that also occurred in 7 of 9 runs also serves to decrease the effective learning rate during ANN training. However, there was a strong tendency to push back (increase) the starting round of learning rate decay, which again occurred in 7 out of 9 runs. This result is surprising, since it seems to contradict the general tendency to decrease the learning rate.

No convincing patterns were observed in the other two parameters, momentum and weight decay. The average momentum term increased in 3 and decreased in 4 runs, with rather strong ($10^{-3} > p > 10^{-6}$) decreases in both members of the Pr and Nr categories. The WtDec parameter changed in all 9 runs, but was split between 4 increases and 5 decreases. There was complete contrast between the Er and Ps categories, with increases in Er2 and Er4, but decreases in Ps2 and Ps4. While the performance of this parameter might interact with the WtDecStart parameter, little can be inferred since WtDecStart showed little to no change.

The ten GA-specific parameters are summarized in Figure 3-9. Two parameters, NumEvo (constrained between 1.0 and 1.99, therefore always rounded down to 1) and MigRate (migration between Evolvables, does nothing when only a single Evolvable exists) were completely unable to affect ANN performance. Changes were observed in

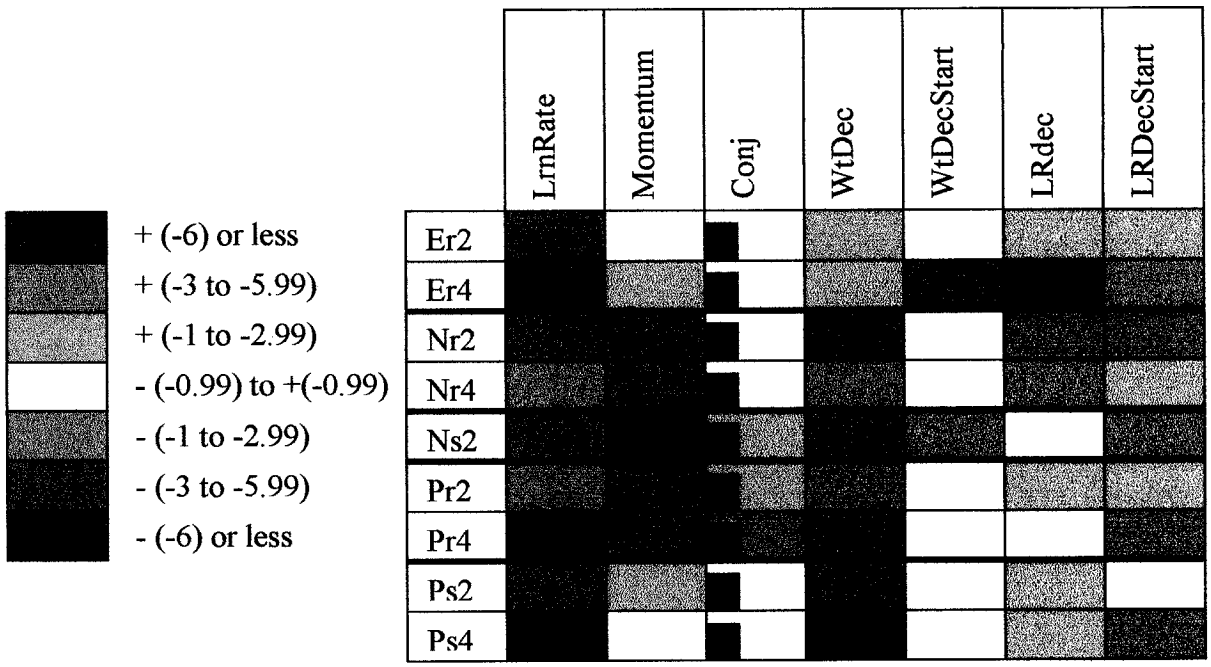


Figure 3-8: Statistical significance of change in parameters optimized by OGA (BP specific parameters). Colour coding is identical to that in Figure 3-7. The GANN runs are identified by type (E = experimental, N = negative control, P = positive control), sequence used (r = real, s = shuffled), and whether input vector optimization was used (2 = no, 4 = yes).

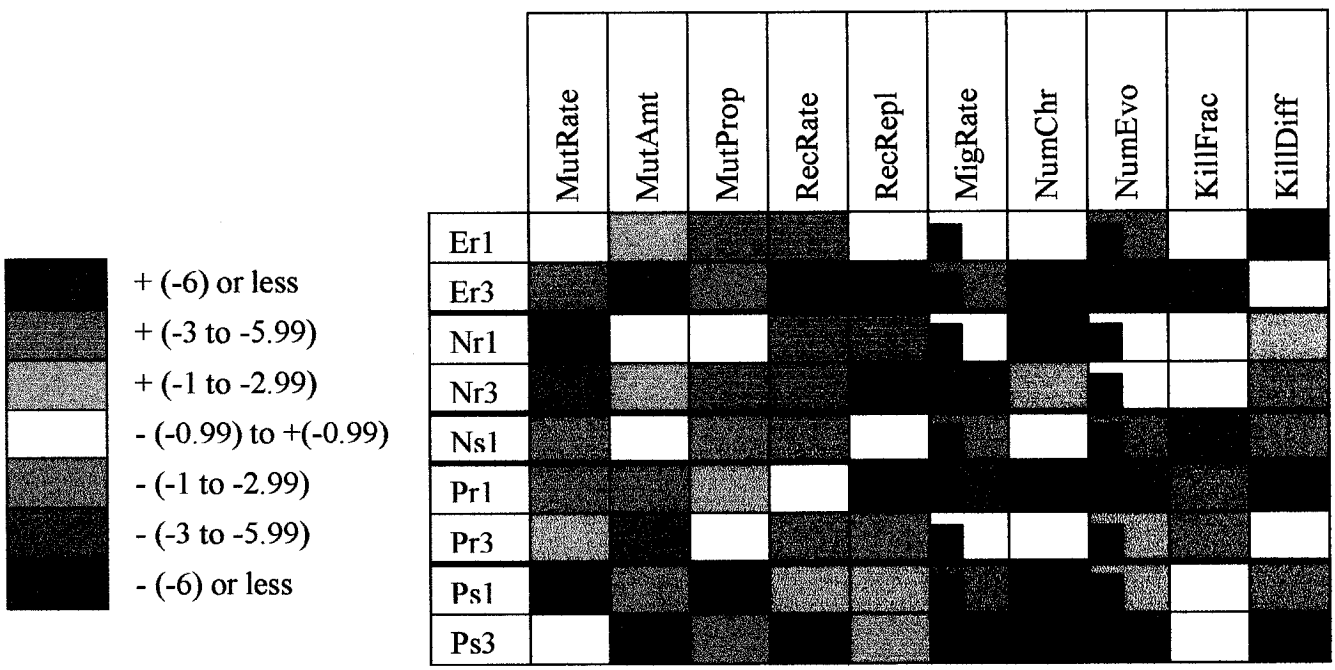


Figure 3-9: Statistical significance of change in parameters optimized by OGA (GA specific parameters). Colour coding is identical to that in Figure 3-7 and Figure 3-8. The GANN runs are identified by type (E = experimental, N = negative control, P = positive control), sequence used (r = real, s = shuffled), and whether input vector optimization was used (1 = no, 3 = yes).

these two parameters in more than half of all runs: NumEvo changed in 7/9 runs while MigRate changed in 6/9 runs, with 4 out of 7 in each category having an associated p-value $< 10^{-3}$. However, no direction of change was preferred.

Several parameters showed consistent patterns of change over most of the nine GA runs. The values of both RecRate and RecRepl increased in a majority of runs, and the only decreases seen for either of these parameters was for RecRepl in the Pr runs, where the relative ease of solving the classification problem might decrease the importance of these parameters. The number of IGA Chromosomes increased in 6 of 9 runs, with an associated p-value $< 10^{-6}$ in five of these.

The three parameters affecting the mutation of connection weights changed in a majority of all runs, but the direction of change was not consistent. Both MutRate and MutAmt increased in four runs and decreased in three, while MutProp increased in three runs and decreased in four. There is no clear evidence of interactions between the terms, since an increase in one of the three terms does not predict an increase in either of the two others. The strongest consistent pattern is the decrease of all three parameters in the 'Ps' runs.

The factors affecting the elimination of similar IGA Chromosomes within the population also showed little consistency. The fraction of Chromosomes considered for elimination did not change in 5 out of 9 runs, and showed no consistency within any of the groups. The KillDiff parameter, which determines the degree of difference below which Chromosomes are considered to be identical, decreased in 5 out of 9 runs. This effect may indicate an advantage to the preservation of very similar Chromosomes, but most of the observed changes were small ($p < 10^{-3}$).

Models of Generalization Accuracy

As another method of assessing which parameters are important to generalization, General Linear Models (GLMs) were constructed. The GLM combines ANOVA-style analyses of categorical variables with regression-style analyses of continuous variables to build an overall model that shows the impact of all contributors to the dependent variable. GLM analysis was performed separately on the entire set of OGA Chromosomes over 30

training rounds for all 18 runs. Using a reverse model of variable inclusion, which first considers all possible contributors but eliminates those that contribute little to the overall model, a final model can be constructed that only includes important contributors to the generalization accuracy.

In addition of the inherent assumptions of the *t*-test described above, collinearity of the 'independent' variables must be considered when constructing a GLM. If correlations exist between different variables, then the final model may choose either or both without reflecting which is the important contributor. In the present analysis, this effect is likely to occur because of the 'hitchhiking' effect, where certain values of less important parameters are preferentially selected due to their association with good values of important parameters. However, in examining the models generated by GLM and the correlations between variables, it may be possible to determine which variables are consistent contributors to most or all models.

Figure 3-10 shows the models constructed using the reverse method of variable exclusion, with a *p*-value of less than 0.001 required for retention of a parameter within the model. The most consistent parameters in these models are those that showed the most consistency in change over time in the *t*-test analyses. Within the shared parameters, the window size decreases while the number of hidden nodes increases. The learning rate shows a consistent decrease among most BP runs, while the start of LR decay is usually pushed back to a later point in the training phase. Within the GA specific parameters, the recombination rate tends to increase, as does the number of IGA Chromosomes.

Conclusions

Experimental Runs Do Not Outperform Negative Controls

Since the goal of GANN is to learn the important patterns in upstream sequence, and to use knowledge of these patterns to classify previously unseen cases, the set of experiments described in this chapter were unsuccessful in achieving that goal. The generalization scores achieved were similar for most of the runs, including the

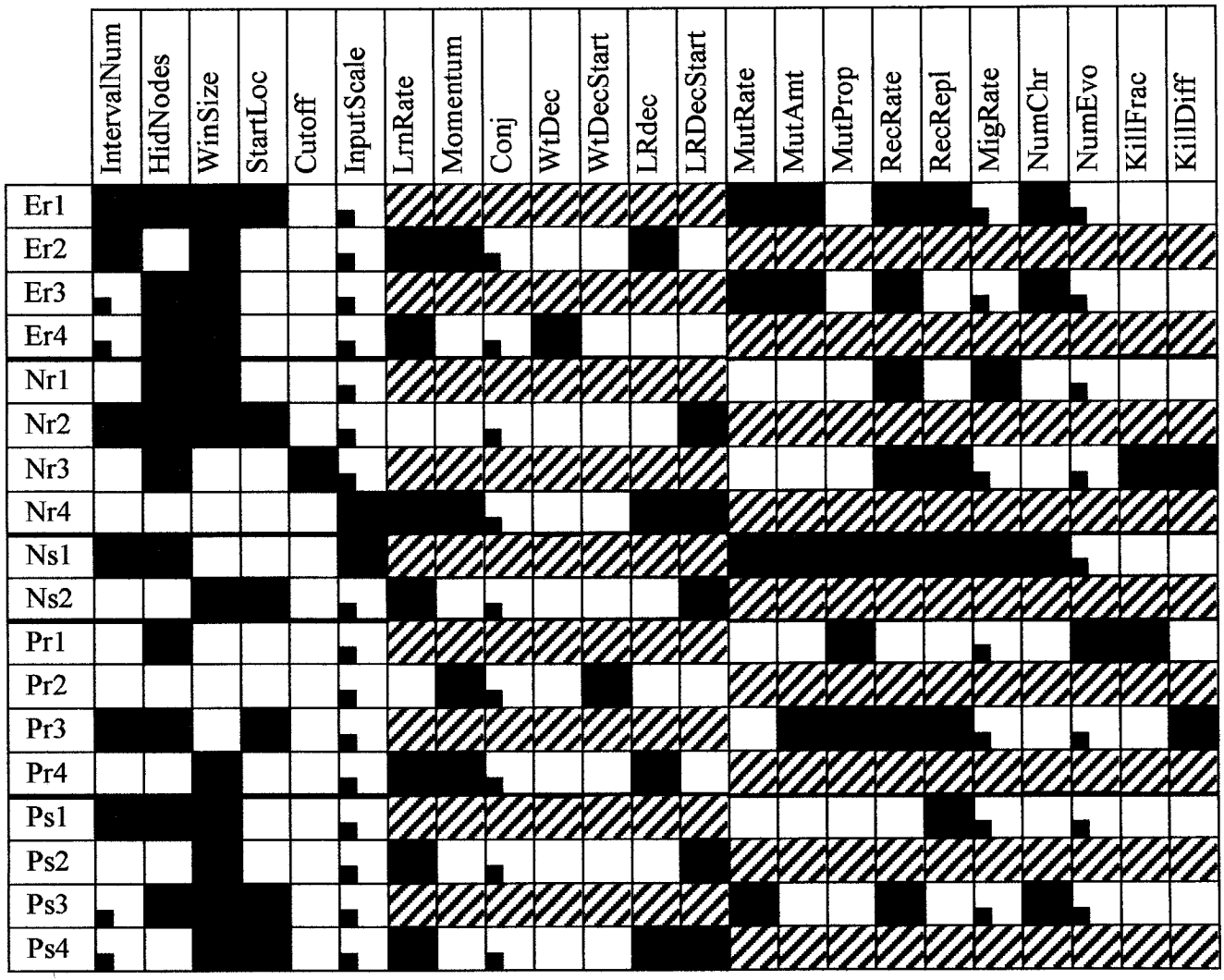


Figure 3-10: Parameters retained by a reverse (exclusion) method in a General Linear Model. The sign of the parameter's contribution to the model is indicated with a red (+) or blue (-) square. A hatched box indicates a parameter this is not applicable to the run performed, while an inset black square indicates a parameter that had no effect on the model. The GANN runs are identified by type (E = experimental, N = negative control, P = positive control), sequence used (r = real, s = shuffled), learning algorithm (1,3 = genetic algorithm; 2, 4 = backpropagation) and whether input vector optimization was used (1, 2 = no; 3, 4 = yes).

experimental runs, the runs using shuffled sequence 'spiked' with known promoters, and the negative controls. While there may have been statistically significant differences between some of these runs, the generalization scores rarely exceeded 0.6 and were therefore not useful for predictive purposes. Since the classification accuracy of real data did not exceed that of shuffled or randomized data, there is no evidence that any useful patterns were learned from the promoter or upstream sequences.

The neural networks created by GANN were often able to learn and classify the training set with a high degree of accuracy, even in some cases where no underlying patterns were assumed to separate the positive from the negative set. This effect was most pronounced when supervised learning was used to train the ANNs, and training set scores in excess of 0.9 were often observed. Since these strong training scores did not translate into good generalization, it is likely that the ANNs were too large and had too many degrees of freedom, as represented by the large number of connection weights. The GA showed a lesser tendency toward very high training set scores, and the generalization scores it produced were equal to or better than those obtained with BP.

The performance of the 'easy' positive controls showed that both the GA and BP methods can theoretically learn a mapping rule if signals are present in the data. While the positive controls used here would be easily solved by any basic statistical classification method as well, they at least demonstrate that GANN can handle certain types of problem. The key to improving the detection and classification of patterns lies in finding a suitable method for encoding the data, and in optimizing the signal-to-noise ratio within those data.

What Parameters Are Optimized By the System?

While the detection and classification of patterns was not good, GANN was able to optimize the ANN architecture and parameters to yield improved generalization. A few key parameters were consistently changed within the population of OGA Chromosomes. The number of hidden nodes increased in almost every run, with the final population mean often very close to the permitted maximum of 50. This finding is somewhat surprising, since larger networks typically tend to memorize the training data without

learning the important underlying rules (Hassoun, 1995). An ANN with 112 input and 50 hidden nodes has a total of (112 x 50) + (50 x 1) = 5650 connections to optimize, which is far too large in relation to the size of the training set. Two actions can be performed to limit the number of connection weights: reducing the upper bound on the number of hidden nodes, and using input vector optimization to optimize the choice of input nodes.

The decrease in mean window size was often dramatic, from an initial value of ~130 to a final value between 60 and 70. This effect may hint at another problem with the current encoding: since the detection method must consider at least 50 nucleotides in a single window, important motif signals may be overwhelmed by the surrounding intergenic sequence. The preference for shorter windows may indicate an attempt by the system to optimize the signal-to-noise ratio.

The other important parameters affected the rate of change of ANN connection weights. The two BP parameters showing the strongest directed change were the learning rate and the learning decay rate, with changes to both parameters leading to highly significant drops in the overall learning rate. This tendency leads to smaller incremental changes in the ANN connection weights, which may lead to better stability in the long run. A learning rate that is too high may bias the network too strongly in favour of the case most recently learned, and may be less likely to lead to convergence.

The two most important GA parameters appear to be the recombination rate and the number of IGA Chromosomes. The tendency toward a larger recombination rate meant that a larger fraction of parents were able to contribute to the next generation, thus preserving a larger degree of genetic diversity within the population. The effect of increasing the number of IGA Chromosomes was similar: with a larger pool of available weights for each connection, the overall diversity of the population was much higher.

Considerations For Future Analyses

As mentioned above, several key factors must be considered in the next analysis. With respect to pattern recognition in general, the large networks generated in this analysis are impractical both in terms of computation and of overfitting. The important quantity is the ratio of training cases to the number of connection weights in the ANN. To

achieve a better ratio with the current data set, the number of input and hidden nodes must be reduced dramatically, and the number of training cases increased if possible (Kasabov, 1996).

Another important issue is the signal-to-noise ratio. Since motifs are not evenly spaced within upstream sequences, the ANN cannot be directed to the exact location of a given motif within each upstream sequence that contains it. However, if the motifs are found preferentially within a restricted range of the upstream region, then shrinking the size of the sequence window that is converted to input indices will reduce the amount of uninformative intergenic sequence. An extreme example of this type of motif is the Shine-Dalgarno sequence, which is usually found within the first 15 nt upstream of the start codon of *E. coli* genes (Fuglsang and Engberg, 2003). While partitioning the upstream sequence into windows will increase the number of input indices in the pool, OGA can be used to search among these indices for those that are the most informative.

Finally, depending on an interval representation of DNA alone may not be adequate for pattern detection. The advantage of this type of measure is its independence from the need for aligned sequences. Interval representations of a similar type were exploited by Ponomarenko *et al.* (1999*b*) to generate novel representations of DNA sequence. However, structural features of DNA and short, possibly degenerate sequence words should be considered as well.

While the experiments described in this chapter failed to construct accurate models of upstream regions, they illustrated some of the important properties of GANN. A few variables seemed to drive the evolution of OGA Chromosomes, depending on the learning method used. Input vector optimization reduced the number of indices presented to the ANNs without decreasing the prediction accuracy and permitted the use of smaller networks. While changes to the input representation must be made, future experiments should exploit the ability of the OGA to choose among different representations of DNA sequence and structure.

[4]

Analysis of Upstream Sequences in *Escherichia coli* K12 and *Sulfolobus solfataricus* P2 Using GANN

Motivation

The experiments performed in Chapter 3 yielded prediction scores that were slightly better for the experimental runs than for the negative controls, but not sufficiently better to demonstrate the learning of useful patterns. This low prediction accuracy may be due to limitations in the detection algorithm, the encoding of the sequence data, or in the original data themselves. Since good predictions were achieved with the positive controls, the core of the detection algorithm (neural networks trained with BP or GA) is here maintained. However, the extraction and encoding of genomic data must be reconsidered, and signal-to-noise optimization must be performed.

The goal of the experiments described in this chapter is identical to that of Chapter 3: attempt to identify patterns that can be used to characterize upstream sequences. While the previous set of experiments produced indices that did not rely on any alignment of the original sequences, trying to find short patterns within a window of size 150 nt or larger with this approach may be impossible due to the noise generated by surrounding sequences. To address this problem, a sliding window approach is used to permit the detection of patterns that may be confined to a part of the upstream region.

Since sequence intervals alone were not sufficient for pattern detection in the last experiment, the new approach uses several new methods to encode DNA sequence. A reduced set of intervals is used, but complemented with the use of short words of DNA (*n*-mers), as well as dinucleotide-based representations of DNA structure. These features alter the relationships between different oligonucleotides, such that two different DNA words with different sequences may still show similar structural features. This property was vital in the analysis performed by Steffen *et al.* (2002), who showed that IHF binding

sites in *E. coli* have very little sequence conservation, but all tend to have extremely low values of deformation energy.

With the implementation of windowing methods and an expanded set of indices, the requirement for an algorithm to choose the best indices is increased. In these experiments, the OGA is used to optimize the inputs to a population of ANNs as before, but the number of inputs is kept constant and is much smaller than in the previous experiments. This constraint should reduce the number of indices that 'hitchhike' within the population, since they must compete with informative indices for inclusion in successive generations of ANNs. Since they all derive from combinations of the four nucleotides, many of the indices used will also be highly correlated with one another. However, if the same useful information is conveyed by two similar indices, there might be no selective advantage to keeping both within a single OGA Chromosome. In the restricted design proposed here, fitter Chromosomes may result if other indices that convey important information take the place of one of the highly correlated pair of indices.

To support the use of GANN as a pattern detection and classification tool, the predictions obtained with this method were compared to those obtained with linear discriminant function analysis (DFA) (Zhang, 2000). DFA is a deterministic statistical classification method that projects the input data into multidimensional space, then generates a hyperplane in that space that maximizes the separation of points belonging to the positive set from points belonging to the negative set. By definition, DFA can only accurately classify problems that are linearly separable, so its ability to learn and generalize should be equal to or worse than that of GANN.

Experimental Design

Sequence Extraction

Two categories of DNA sequence were extracted from different positions within the genomes:

The POSITIVE SET consisted of sequences that were expected to contain regulatory motifs. These sequences were defined as the 150 nt of upstream sequence closest to the start codon of each open reading frame (ORF). Upstream sequences that overlapped another ORF were excluded from the set, which eliminated densely packed genes, including non-leading genes within operons, from consideration.

The NEGATIVE SET consisted of 'other' intergenic sequences. These 150-nt sequences were extracted at random from any genomic DNA that was noncoding and not already in the positive set.

Sequence Splitting

The 150-nt sequences contained in the positive and negative sets were then subdivided into smaller 'windows' of sequence. Window sizes of 10, 25, 50 and 75 nt were considered, with an arbitrary 40% overlap between adjacent windows. Since the overlapping windows did not always cover exactly 150 nt of sequence, some window sizes excluded bases at the upstream end of the sequence. The windows were assigned numbers to indicate their position within the 150 nt sequence, with the window closest to the start codon identified as 0 and the most upstream window assigned the highest number.

Generation of Indices

The windows of sequence were each converted to a series of numeric values that represent sequence composition and structural parameters. The indices used in this study are summarized in Table 4-1. For each sequence and structural parameter, values representing the maximum, minimum and mean value for all windows of a given size from a given sequence were also calculated.

Table 4-1: Sequence and structural representations of DNA, with experimental group subdivisions. The indices were subdivided into three groups: the BASIC group contained all 13 structural representations of DNA, as well as the 'nuc-freq' (10 indices) and 'dinuc-freq' (16 indices) categories; the TRINUC group contained all 64 trinucleotide indices; and the NUC_INT group contained 32 nucleotide interval indices. References for the structural conversion rules are shown in the last column.

Variable Group	Description	Index Type	Reference
BrunakStack	Stacking energy	BASIC	Baldi et al., 1998
BrunakProp	Propeller twist: twisting of bases in a pair in opposite directions	BASIC	Baldi et al., 1998
LisserFlex	Helical flexibility	BASIC	Lisser and Margalit, 1994
LisserB→ A	Energy required to change adjacent base-pairs from B helical conformation to A	BASIC	Lisser and Margalit, 1994
LisserB→ Z	Energy required to change adjacent base-pairs from B helical conformation to Z	BASIC	Lisser and Margalit, 1994
OlsonFlex	Another measure of flexibility	BASIC	Olson et al., 1998
FlexAbund	Same measure with reinforcement for adjacent highly flexible base-pairs	BASIC	Olson et al., 1998
Twist	Z-axis rotation between base-pairs	BASIC	Gorin et al., 1995
Roll	Y-axis rotation between base-pairs	BASIC	Gorin et al., 1995
Tilt	X-axis rotation between base-pairs	BASIC	Gorin et al., 1995
Rise	Z-axis translation between base-pairs	BASIC	Gorin et al., 1995
Slide	Y-axis translation between base-pairs	BASIC	Gorin et al., 1995
Shift	X-axis translation between base-pairs	BASIC	Gorin et al., 1995
Nuc-freq	Frequency of each nucleotide (A,C,G,T) as well as IUPAC degenerate bases, with a bonus for adjacent nucleotides that are identical	BASIC	-
Dinuc-freq	Frequency of dinucleotides (AA…TT), with a bonus for adjacent dinucleotides that are identical	BASIC	-
Trinuc-freq	Frequency of trinucleotides (AAA…TTT)	TRINUC	-
Nuc-int	Frequency of dinucleotides with unspecified intervening nucleotides (AA, ANA, ANNA, etc.)	NUC_INT	-

Finally, a score was computed for each index value, by comparing the calculated value with the mean index value calculated from 100 permutations of the window sequence. This quantity was represented by a Z-score:

Z = (calculated value - mean of randomized values) / (standard deviation of randomized values)

(4.1)

OGA Training

The OGA was used to test and select combinations of the thousands of generated indices. A restricted set of ANN parameters were also optimized using the OGA. As in Chapter 3, the fitness of each OGA Chromosome was proportional to the generalization ability of neural networks generated with the specified parameters and trained with the specified indices. The parameters optimized by the OGA are shown in Table 4-2. Since backpropagation-trained ANNs were not used in this analysis, parameters relevant to BP were not optimized by GANN. Some of the parameters optimized in the previous set of experiments (see Chapter 3) were assigned fixed values for the present analysis. Inputs were always standardized, and the number of IGA Chromosomes was fixed at 50. Ten rounds of OGA training were performed.

Table 4-2: Summary of parameters optimized by the OGA, with maximum and minimum values. Only a single architectural parameter (the number of hidden nodes) was considered for optimization, with the rest all pertaining to the conditions imposed on GA learning. The parameter abbreviations used in the text are shown, followed by the maximum and minimum values for each parameter.

Parameter Type	Parameter	Abbreviation	Minimum Value	Maximum Value
Architecture	Number of Hidden Nodes	HidNodes	5	15
GA Learning	Fraction of population subject to mutation	MutProp	0.10	0.90
GA Learning	Rate of mutation within chromosomes	MutRate	0.10	0.90
GA Learning	Maximum magnitude of mutation	MutAmt	1.1	1.5
GA Learning	Proportion of parents that recombine	RecRate	0.10	0.90
GA Learning	Proportion of population replaced by recombinants	RecRepl	0.50	0.9
GA Learning	Proportion of Chromosomes subject to similarity-based elimination	KillFrac	0.10	0.90
GA Learning	Maximum Euclidean distance between 'identical' Chromosomes	KillDiff	0.10	0.90

Results

Data Sets and Controls

Extraction of positive and negative sets yielded 611 upstream and 991 non-upstream sequences of length 150 nt from *E. coli* K12, and 927 upstream and 729 non-upstream sequences of the same length from *S. solfataricus* P2. The size of the positive set extracted from both genomes is much smaller than the number of identified ORFs because high ORF density and operon organization result in many upstream regions of length << 150 nt, especially in *E. coli*.

Two types of control were performed in parallel with the experimental data runs:

- Negative control sets were generated by randomly reassigning positive and negative set sequences. The size of each set was not changed, but any patterns initially overrepresented in either the positive set or the negative set would then be spread among the two sets.
- Positive control sequences were created by randomly generating DNA sequences with 60% (A+T), which corresponds to the average composition of intergenic sequence in *E. coli* K12, then 'spiking' the positive set sequences with the σ^{70} promoter consensus TTGACA(N_{17})TATAAT. This procedure was repeated twice, the first time with consistent positioning and spacing of the conserved sequences, the second time with positioning +/- 30 nt, and spacing +/- 1 nt to simulate the natural variability in these quantities while still keeping the entire conserved motif within the 150 nt sequence.

Pre-screening of Indices

The large number of windows, as well as the large number of sequence and structural properties that were considered, yielded 14 417 distinct indices associated with each 150-nt sequence. The number of indices in the NUC_INT and TRINUC categories

was very large, so these two groups were processed and recorded separately from the rest of the indices, which yielded three index types (Table 4-1). Each index was used as the single input to train at least three separate populations of ANNs. Figure 4-1 shows the range, mean and standard deviation of the generalization scores for all indices, with the index scores grouped by window size and index type. While the *mean* scores associated with the experimental trials tend to be only 0.005 – 0.01 higher than the corresponding scores for the negative control runs, the *best-scoring* indices are substantially higher, with most categories exhibiting a difference between 0.05 and 0.1. These differences suggest that the majority of indices in the experimental runs are poor predictors of sequence type, with predictive ability similar to the negative control set. There is a slight upward trend in the mean generalization score with increasing window size, but the maximum scores tend to decrease. This trend is likely due to a better signal-to-noise ratio that would occur in a small fraction of windows of size 10, which would tend to be diluted in larger windows but present in a larger proportion of them. The indices generated from the first positive control sequences (PosA) had a slightly higher mean, but the majority of indices were still uninformative due to the concentration of signal in only a few windows. The best single index score was much higher than any obtained in the experimental runs. The second positive control (PosB), where the position of the conserved sequences were varied, showed a pattern of scores that was very similar to the experimental runs.

To select the best indices, a score cutoff was established for each window size and index group. The cutoff was defined as a value that was greater than 95% of the scores in each experimental group's corresponding negative control run. This cutoff score ranged between 0.505 and 0.52 for different sets. Use of these cutoff values reduced the size of the *E. coli* K12 set to 2053 indices, and the *S. solfataricus* P2 set to 2002 indices.

Ten-input ANN Runs (Separated by Index Type)

The most successful indices of all window sizes from the single-input runs were combined to yield three pools of indices per genome, one for each index type (BASIC, NUC-INT, and TRINUC). A population of OGA Chromosomes was generated for each pool of indices, with the number of Chromosomes approximating the total number of

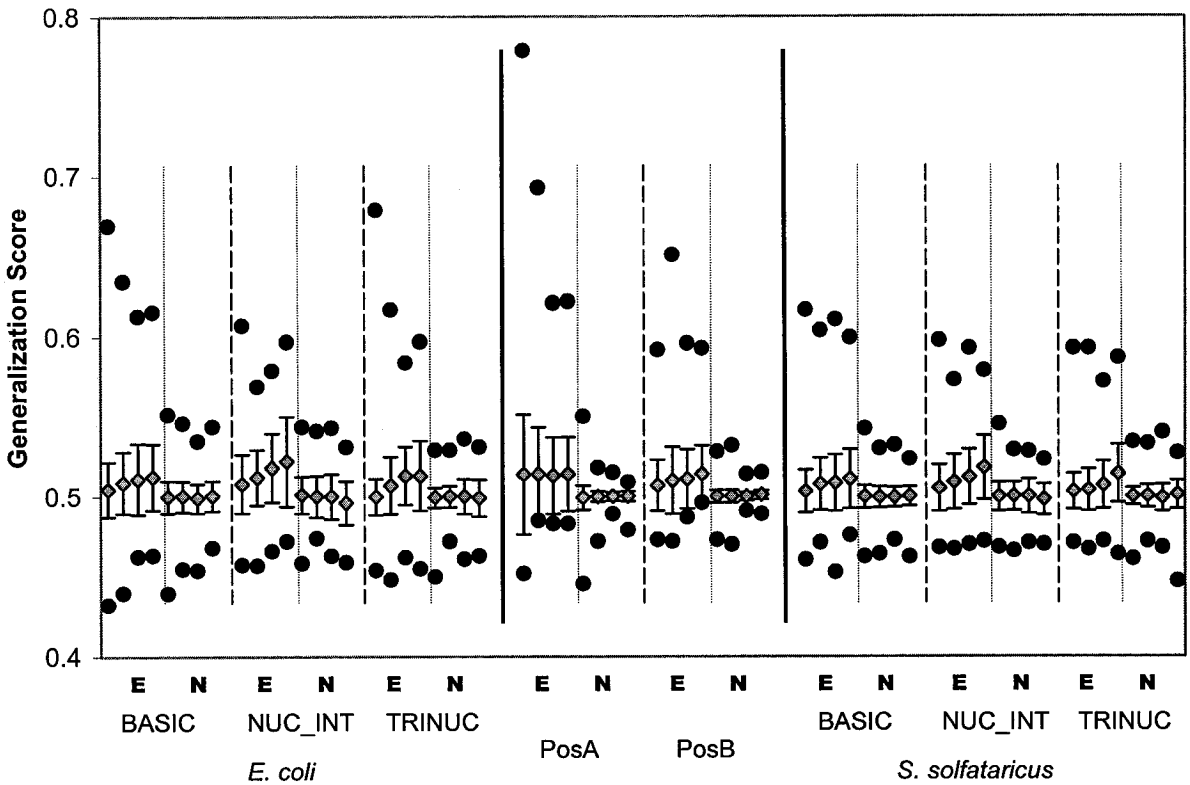


Figure 4-1: Generalization scores, expressed as the difference between true positives and false positives, for single-input runs associated with *Escherichia coli* K12, two positive controls (Pos), and *Sulfolobus solfataricus* P2. The indices are separated by index type (see Table 4-1), and are further subdivided into experimental (E) and negative control (N) runs. Within each grouping, the four window sizes considered (10, 25, 50, and 75 nt) are shown from left to right as a mean score +/- standard deviation. Maximum and minimum values are indicated by circles above and below, respectively.

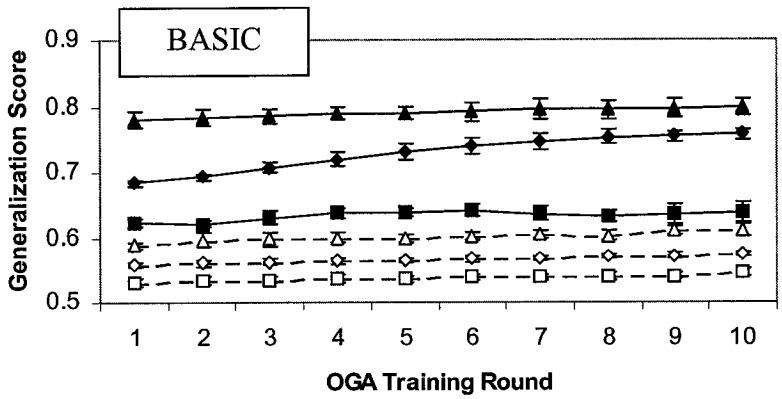
indices. Each OGA Chromosome specified ten randomly selected indices from the pool to be used as ANN inputs, so each index was represented 10 times on average, and 99% of the indices were represented at least three times, based on the Poisson distribution. The ANN populations specified by these OGA Chromosomes were trained and tested as described in the **Experimental Design** section above, and ten rounds of OGA recombination were carried out. This procedure was replicated five times each for *E. coli* K12 and *S. solfataricus* P2.

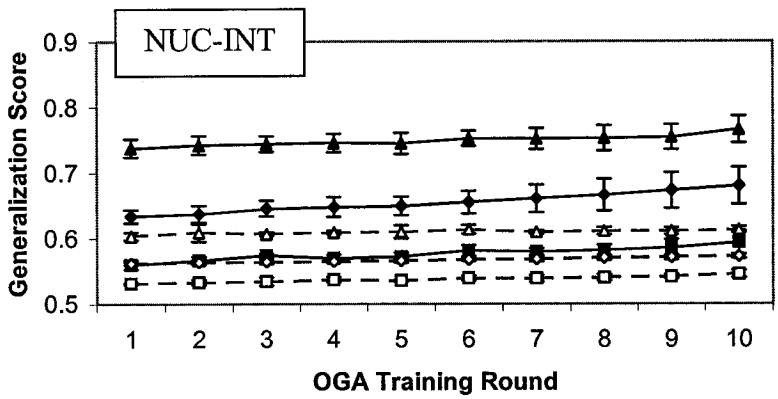
Figure 4-2 shows the improvement in generalization score by comparing the maximum, mean, and minimum OGA Chromosome generalization scores over the ten OGA training rounds for the three different index type pools. These scores are compared with another set of negative controls that were generated by randomizing the set memberships of the pooled indices. In most cases, the score of the best Chromosome in an OGA population improved by only 0.02 – 0.03 over the ten training rounds, while the average score of all OGA Chromosomes improved by between 0.05 and 0.1. This difference in score improvement suggests that the OGA is better at eliminating poor combinations of indices, rather than generating better combinations of indices through recombination. The negative controls performed substantially worse than their experimental counterparts, with only occasional overlap in score between the best negative control OGA Chromosomes and the worst experimental runs. The best OGA scores for the first positive control (PosA) were nearly perfect, and indeed some of the best neural networks within the best OGA Chromosomes achieved a classification accuracy of 1.0. In contrast, the scores for the second positive control (PosB) were similar to those from the *E. coli* K12 runs, increasing to almost 0.8 by the end of training.

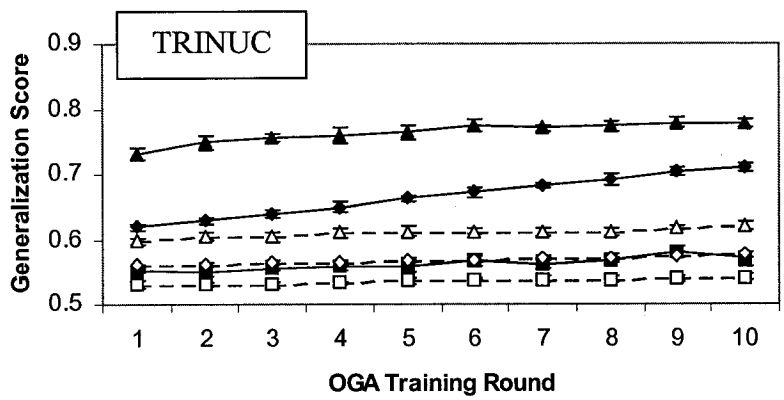
Since the indices were not evaluated independently in this step, a different strategy was required to choose the best indices for subsequent analysis. The indices were assigned a rank based on their average frequency in all five experimental populations after ten rounds of recombination were carried out. This ranked list was compared with a similar ranked list of all negative control indices. The cutoff for acceptance of experimental indices was defined as the rank in the two lists where the frequencies in both ranked lists were most similar. Every index in the experimental list above this cutoff point was potentially subject to positive selection by the OGA, and was therefore

Figure 4-2: Improvement in generalization scores over ten rounds of OGA training, for *E. coli* (*a*), the positive control sets (*b*), and *S. solfataricus* (*c*). The maximum (triangles), mean (diamonds), and minimum (squares) OGA Chromosome scores are shown with standard deviations (n = 5). Experimental scores are depicted with solid symbols and solid lines, whereas the negative controls are shown with empty symbols and dashed lines.

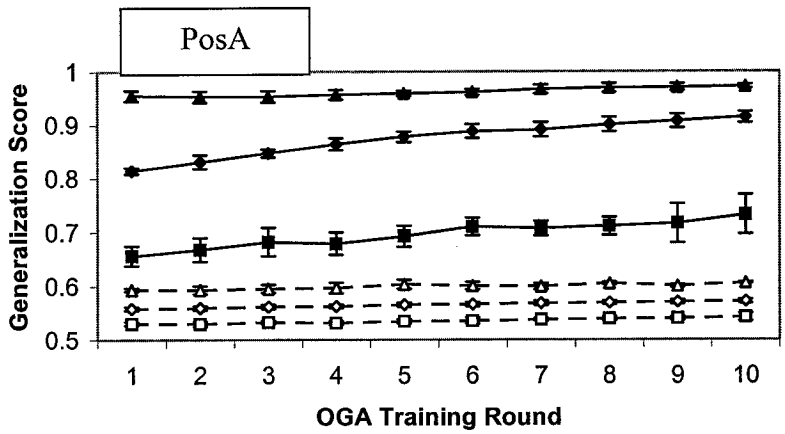
a) *E. coli* K12

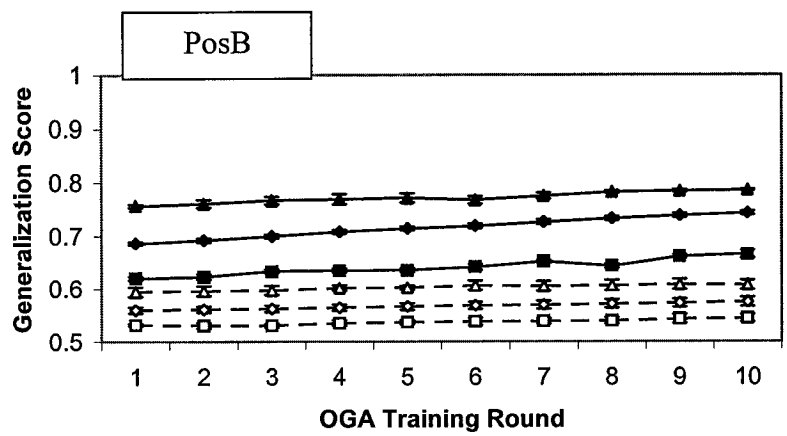

BASIC
Generalization Score
0.9
0.8
0.7
0.6
0.5
1 2 3 4 5 6 7 8 9 10
OGA Training Round


NUC-INT
Generalization Score
0.9
0.8
0.7
0.6
0.5
1 2 3 4 5 6 7 8 9 10
OGA Training Round

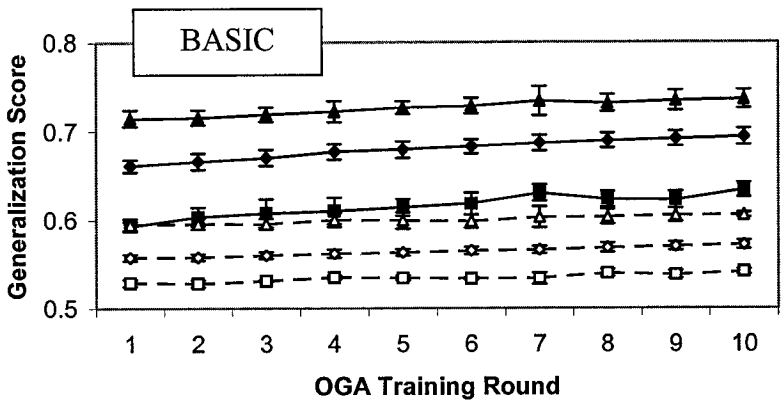

TRINUC
Generalization Score
0.9
0.8
0.7
0.6
0.5
1 2 3 4 5 6 7 8 9 10
OGA Training Round

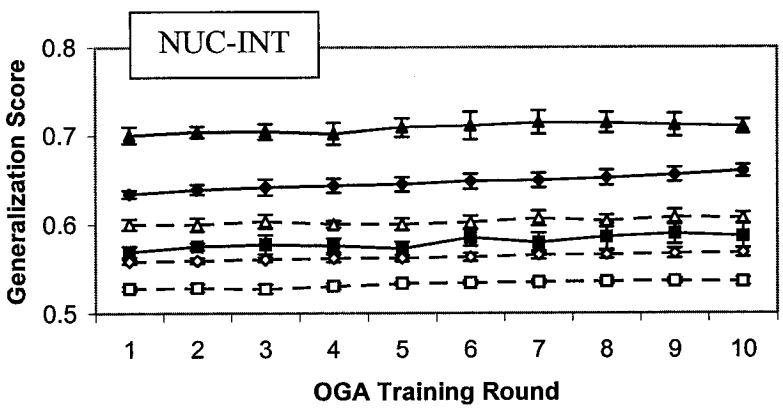
b) Positive controls

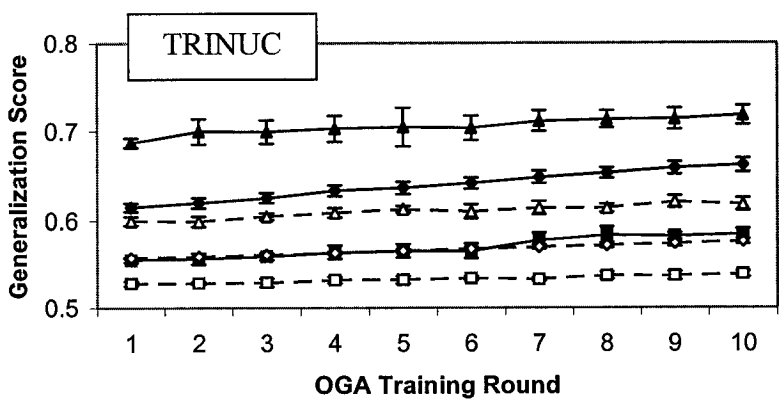




c) *S. solfataricus* P2







included in the final analysis. Using this cutoff, the number of indices for *E. coli* K12 and *S. solfataricus* P2 were reduced to 198 and 309, respectively.

Varying the Number of Input Indices

The successful indices from all index types in the previous analysis were combined into a single list for each genome in the final set of experiments. Five separate OGA trials were conducted for each set of indices, with either 2, 4, 8, 16 or 32 indices at a time presented to the ANN. The best neural network generalization score from each of these runs is shown in Figure 4-3. There appears to be little improvement in generalization scores for *E. coli* K12 above 8 inputs, but *S. solfataricus* P2 scores do not reach a plateau, even at 32 indices. Given the lack of independence between indices, however, it is unlikely that using more than 32 indices will yield any improvement in generalization.

Pattern Localization and Association

Table 4-3 shows the distribution of the 50 most successful indices along the 150 nt upstream regions of *E. coli* K12 and *S. solfataricus* P2, respectively.

The strongest pattern that emerged from this analysis is located in the region directly adjacent to the start codon in *E. coli*. A number of indices from all window sizes were strongly represented in the final population, and the preponderance of indices relating to purine frequency suggests that the Shine-Dalgarno ribosome binding site (RBS; Shine and Dalgarno, 1974) is being identified. Although the sequence itself is small and a window size of 10 is clearly best for detecting this signal, the RBS is still detected in window sizes of 25 and greater. Other signals may have been detected as well, but not obviously. In particular, the canonical σ^{70} promoter sequence, TTGACA(N_{17})TATAAT, was not clearly identified.

A similar type of pattern was detected adjacent to the start codon in *S. solfataricus*, which likely corresponds to the archaeal equivalent to the Shine-Dalgarno sequence. However, the signal is much weaker, which suggests either that the sequence is

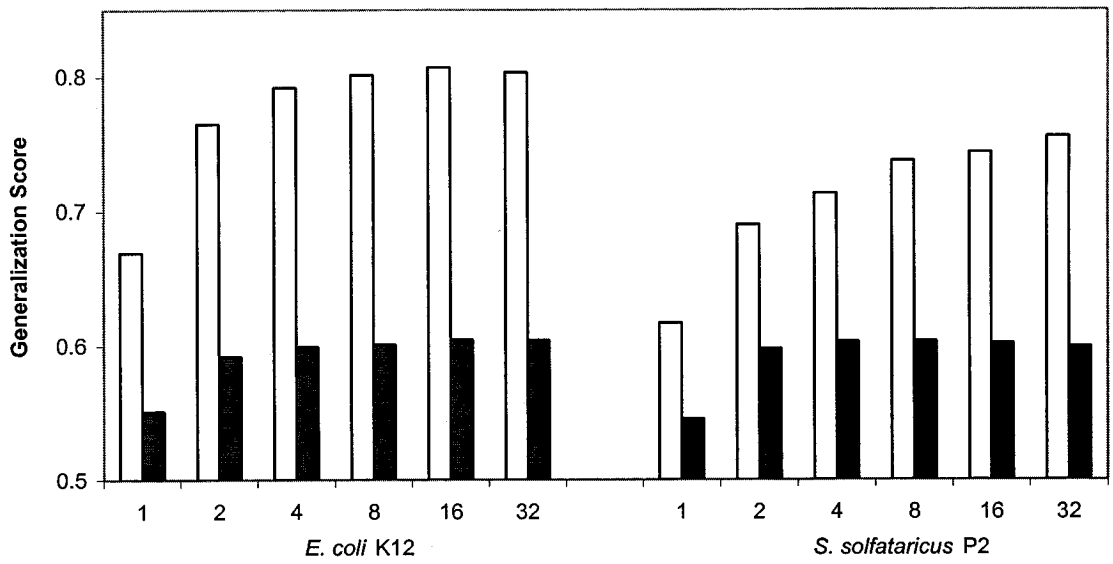


Figure 4-3: Best OGA chromosome generalization score obtained with different numbers of input indices. White bars indicate experimental runs, and negative controls are shown in black.

Table 4-3: The 50 most successful indices from the combined runs of *E. coli* K12 and *S. solfataricus* P2. Within each separate run (of 2, 4, 8, 16 or 32 inputs), each input was ranked on its frequency x in the final population of chromosomes. Each value x was then transformed with the equation $y = \log(x + 1)/\log$ (number of OGA Chromosomes). The final rank or relative 'success' of an index was then the mean of each of its y scores in all five runs. The indices are sorted by window size, and the location of the window within the 150 nt extracted sequence is shown, where '-1' is at the 3′ end of the sequence, and '-150' at the 5′ end. GLOB refers to the windows that represented the maximum, minimum and mean of all window scores for a given index type. The indices here are consistent with those described in Table 4-1: Different nucleotide combinations (including IUPAC degenerate codes) represent mono-, di- and trinucleotide frequencies, a pair of nucleotides separated by a number refer to nucleotide intervals with the specified number of intervening N's, and the suffix '-Z' refers to a Z-score.

Size	Position (window #)	*E. coli*	*S.solfataricus*
10	-1 to -10 (23)	R, Y, GGA, AGG, GGA-Z, AGG-Z, S-Z, GA, G-Z, G, GG-Z, GAG, GAG-Z	GGT-Z
	-7 to -16 (22)	C0G, C0G-Z	
	-13 to -22 (21)		BrunakStack, FlexAbund, LisserB→ Z, Slide
	-19 to -28 (20)	G0G	TTA, BrunakStack, OlsonFlex, S, M-Z, LisserB→ Z, W, TTA-Z, GAA-Z
	-25 to -34 (19)		TTG-Z, K-Z
	-31 to -40 (18)		
	-37 to -46 (17)	G0G	
	-43 to -52 (16)		
	-49 to -58 (15)	T0G-Z	
	-55 to -64 (14)		
	-61 to -70 (13)		
	-67 to -76 (12)		
	-73 to -82 (11)		
	-79 to -88 (10)	GCA	
	-85 to -94 (9)		A1A
	-91 to -100 (8)		
	-97 to -106 (7)		
	-103 to -112 (6)	A0G	
	-109 to -118 (5)		
	-115 to -124 (4)		
	-121 to -130 (3)		
	-127 to -136 (2)		T1A
	-133 to -142 (1)		
	-139 to -148 (0)	A0G	A1A-Z
	GLOB		FlexAbund-MAX
25	-1 to -25 (8)	C0A, GGA-Z, R	Slide, TA, OlsonFlex, BrunakStack, TTA
	-16 to -40 (7)		TTC, TTA
	-31 to -55 (6)	T0A-Z	TTC
	-46 to -70 (5)	A0G	
	-61 to -85 (4)		
	-76 to -100 (3)	GCT-Z	
	-91 to -115 (2)	W, ATT, A0C	
	-106 to -130 (1)		A1A, A0A
	-121 to -145 (0)	AAT, AAT-Z	G0T
	GLOB		T-AVG, Slide-AVG
50	-1 to -50 (3)	GGA-Z, GGA	TC, CGT, TCC, TTT, AAT-Z
	-31 to -80 (2)	T1A, A0C	AAT-Z, G0G
	-61 to -110 (1)	ACC-Z, LisserB→ Z, C1A	
	-91 to -140 (0)		C0C, A0G
	GLOB	FlexAbund-AVG, TT-MAX	TC-MAX
75	-1 to -75 (1)	C1T, G1G, GAG-Z, AG-Z	TA, TTG-Z, C0T
	-46 to -120 (0)	TTT, AG-Z	T0G
	GLOB	FlexAbund-MAX, BrunakStack-MIN, W-MAX	C-MIN, TC-AVG, TA-MIN

more weakly conserved here than in *E. coli*, or that the RBS is not found in all upstream sequences of *S. solfataricus*. This hypothesis is consistent with the observation that Shine-Dalgarno sequences are not required for translation initiation of many archaeal transcripts (Slupska *et al.*, 2001).

Stronger than the Shine-Dalgarno signal in *S. solfataricus*, however, were a group of indices located ~30 nt upstream of the start codon. These indices are consistent with the archaeal TATA box which is the consensus binding site for RNA polymerase (Soppa, 1999). The successful detection of the promoter sequence in *S. solfataricus* P2 where none was clearly identified in *E. coli* K12 may be indicative of more consistent spacing between the transcription and translation start sites, a higher frequency of archaeal TATA boxes in the training set of *S. solfataricus* compared to σ^{70} binding sites in *E. coli*, or a stronger adherence to the consensus in *S. solfataricus*. As with *E. coli*, a number of other indices were selected for in the population that could not be associated with a known regulatory feature.

Optimization of ANN Parameters

While consistency in ANN parameter evolution was assessed in Chapter 3 based on comparison between different runs, the use of replication in the present experiments provides a better opportunity for testing the consistency of parameter evolution. The set considered in this analysis was derived from the second step of input optimization, when unified runs for all BASIC, TRINUC, and NUC_INT (see Table 4-1) indices were performed for *E. coli* K12 and *S. solfataricus* P2. A single replicate from each set of indices was analyzed in detail.

Many of the parameters from Chapter 3 (see Table 3-2) are not relevant to this analysis. Since backpropagation-trained ANNs were not used in this analysis, parameters such as learning rate and momentum were never used. Also, the window size and start location parameters were lost, since these factors were implicit in the step where windows were extracted from the genome prior to GANN analysis. Optimization of input indices using OGA eliminated the IntervalNum parameter. Other parameters were set to a constant value in light of their behaviour in the previous analysis: input scaling was

always performed, the cutoff value was set to 0.0 (with ANN output node values ranging between –1.0 and 1.0), and the number of IGA Chromosomes was fixed at 50.

As in Chapter 3, *t*-tests were performed to measure the statistical significance of differences between parameter values in OGA generation 1 and generation 10. The magnitudes of *p*-values associated with each parameter for each of the runs considered are shown in Figure 4-4. Red squares indicate an increase in mean parameter value with an associated *p*-value less than 0.1, while blue squares indicate a decrease with the same constraint. While the legend is consistent with the similar figures from Chapter 3, the sample size has increased from 25 or 50 in Chapter 3 to 950 in these experiments. This increase is likely to yield more *p*-values that are low (e.g., < 0.001) by chance.

The two most important parameters from the previous analysis, included here, are the number of hidden nodes (HidNodes), and the fraction of parents that are permitted to recombine (RecRate). These two parameters showed a strong directional change in these experiments as well, with RecRate increasing in all 12 runs and HidNodes increasing in 7 out of 12 cases. These changes are consistent between the experimental and negative control runs. The difference in mean parameter values between the first and final generations was large enough to yield very low *p*-values, but the difference in comparison to the allowed range was quite small. While the final value of this parameter was often close to the maximum allowed value in Chapter 3, the change in the mean HidNodes here reflect an average change of less than 1. Even the very small *p*-values associated with the RecRate parameter reflect a change of about 0.05 within an allowed range of 0.1 and 0.9.

The other optimized parameters showed a bias equal to or less than that seen for HidNodes. The RecRepl parameter, which defines the fraction of parental Chromosomes replaced by recombinants, increased in 7 cases and decreased in a single run, a slightly weaker bias than that seen with HidNodes. The KillFrac parameter, which defines the fraction of IGA Chromosomes that can be destroyed due to their similarity to other members of the population, showed little change, with 2 increases and 4 decreases of the mean value in 12 runs. There was a tendency to increase the minimum degree of similarity between a pair of IGA Chromosomes required for elimination of one or the

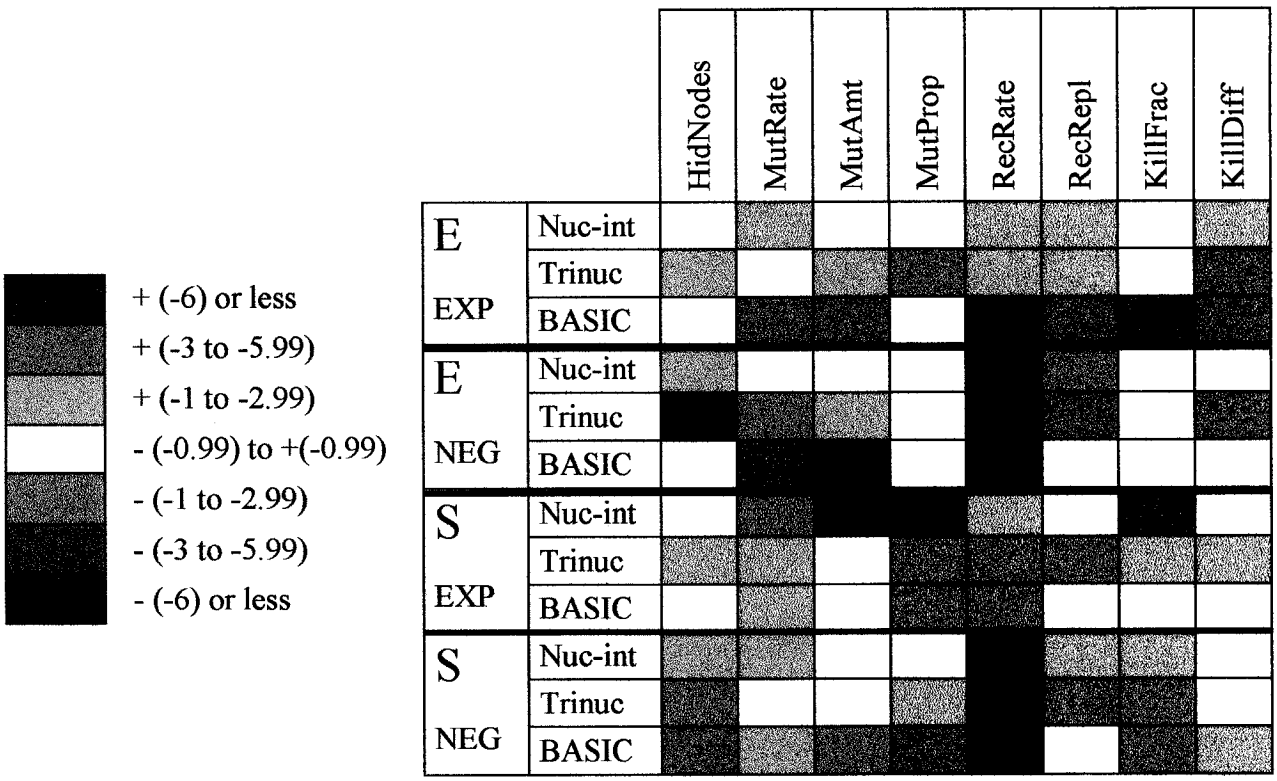


Figure 4-4: Statistical significance of change in parameter mean values over 10 rounds of OGA training. The colour of each box is determined by the base 10 logarithm of the *p*-value associated with the change in the parameter value over 30 OGA generations. The darkest shade of red indicates an increase with a *p*-value of less than 10^{-6}, while the darkest shade of blue indicates a decrease of the same magnitude. GANN runs are subdivided by genome (E = *E. coli* K12, S = *S. solfataricus* P2), type of run (EXP = experimental, NEG = negative control), and by index category (see Table 4-1 for grouping of indices into Nuc-int, Trinuc and BASIC categories).

other (see *Genetic Algorithms (GANet)* in Chapter 2), as the KillDiff parameter increased in 6 out of 12 runs, with associated p-values always $>10^{-6}$. The increase in the number of recombinants generated is consistent with the weak pattern in Figure 3-7, but the increase in KillDiff seen here conflicts with the pattern of that parameter in the same figure. The parameters affecting mutation all showed a tendency to increase through the OGA training rounds, but there is no indication of the strong bias seen for RecRate.

Discriminant Function Analysis

DFA was performed on the same sets of *E. coli* K12 and *S. solfataricus* P2 indices that were used in *Varying the Number of Inputs* above. To assess the generalization ability of trained DFA models, the data sets were randomly split into training and test subsets a total of 10 times. Each of the training sets thus obtained was used to generate a separate discriminant function. A reverse training method was used, with all parameters initially included in the model, but with subsequent removal of all parameters that did not have at least a minimal significance level within the model. All indices were at first included in the model, but indices that did not have an associated p-value < 0.001 were removed in turn. Once generated, the discriminant function was used to classify the test set. Since this method allowed the identification of true positives and false positives, a QA score was calculated in the same manner as for GANN above.

Negative control data sets were created by randomizing the membership of sequences from the positive and negative sets. The ratio of positive to negative sets was kept identical in the experimental and negative control sets for each genome. This randomization was also performed ten times for each genome, and each of these randomized sets was split once into a training and test set. DFA models were trained using a reverse variable selection method as above, but a less stringent p-value cutoff of 0.01 was used. This modified cutoff was necessary because no variables were retained in the negative control models when the maximum p-value was set to 0.001.

The mean generalization score for each set of trials is shown in Figure 4-5. While some variables were always retained in the experimental models, some of the negative control models did not possess any variables with associated p-values < 0.01. Thus, the

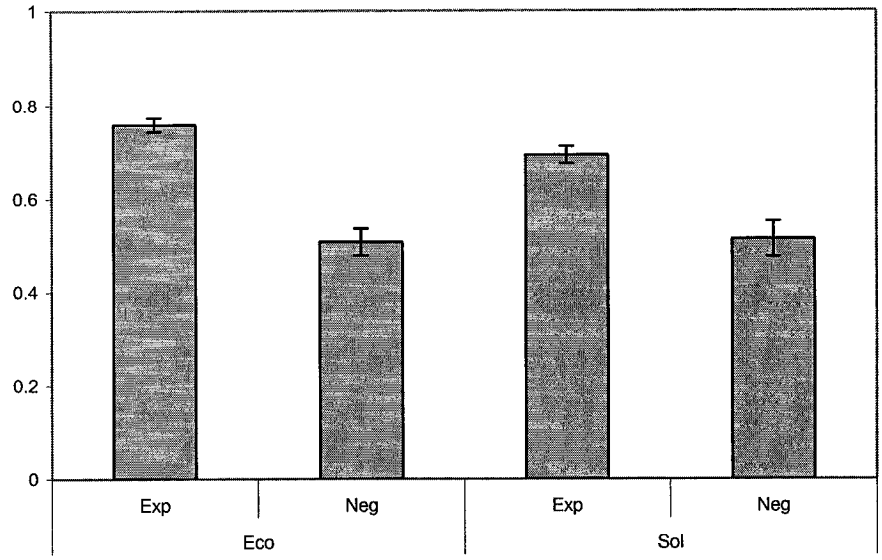


Figure 4-5: Mean generalization scores of models generated with discriminant function analysis (DFA). Standard deviations are indicated, with $n = 10$ for both experimental runs, 5 for the *E. coli* K12 negative control, and 7 for the *S. solfataricus* P2 negative control.

mean and standard deviations of the negative control runs are based on 5 replicates for *E. coli* and 7 for *S. solfataricus*. The DFA models derived from experimental data sets yielded lower generalization scores than the variable input GANN runs shown in Figure 4-3, for any number of GANN inputs greater than 2, but the difference is not dramatic. The mean DFA generalization score for the *E. coli* sets is 0.759, while the GANN generalization scores for 4 to 32 nodes are all between 0.79 and 0.81. The disparity in scores for *S. solfataricus* is slightly less: DFA models yielded an average generalization score of 0.694, while GANN yielded a range of scores between 0.71 for 4 input nodes and 0.755 for 32 input nodes. Since the DFA models generated for *E. coli* retained between 8 and 11 variables, while those for *S. solfataricus* retained between 13 and 17, GANN was able to achieve better generalization scores with fewer input indices.

Conclusions

In these experiments, GANN learned to classify upstream sequences versus non-upstream sequences with 80% accuracy in *Escherichia coli* K12, and 75% accuracy in *Sulfolobus solfataricus* P2. While these levels of accuracy are insufficient for reliable identification and prediction of regulatory features, the system was still able to identify important patterns in both genomes. The evidence that such patterns were found lies in the examination of the inputs selected by the OGA. There was an uneven distribution of strongly selected inputs within the 150 nt considered, and several window positions within this sequence were strongly favoured over others.

Some of the most successful indices could not be linked to any characterized regulatory protein binding site. These signals may represent sequence or structural features that are characteristic of upstream regions, but that do not specifically interact with regulatory proteins. Such patterns are described in the GenomeAtlas database (Pedersen *et al.*, 2000). Some indices may have a higher frequency due to biologically insignificant patterns that are slightly over- or underrepresented in upstream regions. Finally, some indices that are statistically neutral may spread through the OGA population due to genetic drift. The change in index frequency observed in the negative control runs should be due entirely to the last two effects, so these controls are good

estimators of the biological relevance of the experimental runs.

Given the increase in sample size used for statistical analyses, the observed changes in parameters optimized by OGA were less significant here than in the analysis from Chapter 3. Since the OGA was trained for only 10 rounds instead of 30, there is also less time for ANN parameters to show drastic changes in distribution. However, the most important methodological change affecting these parameters was the extensive use of the OGA to optimize the inputs to the ANN. Since the choice of inputs can potentially have a much greater impact on generalization ability than subtle changes in ANN parameters, the input indices are likely to drive the selection of OGA Chromosomes. If a stable population of input indices were eventually obtained, then 'fine-tuning' of the system might then drive important ANN parameters to extreme values. The only convincing change in any ANN parameter was observed for the recombination rate. A strong tendency for increase in this rate suggests that it is advantageous to permit more IGA Chromosomes to contribute their connection weights to the next generation, instead of letting only the very best IGA Chromosomes reproduce.

The best generalization scores obtained in the experimental runs were consistently higher than the best negative control scores (Figure 4-1), with little overlap in the ranges of the two types of runs (Figure 4-2). This difference strongly suggests that the best indices from the experimental set correspond to real patterns. We cannot, however, reconstruct fully functional regulatory sequences from these indices alone, because the indices used in these experiments were not intended to identify complete (or maximal) biological features. Also, it is inherently difficult to determine how specific sets of indices can influence a trained neural network, because of the multitude of connection weights and the non-obvious nature of interactions among them. Still, the indices should represent important components from upstream sequences, and could thus suggest targets for further research.

Analysis with DFA showed that GANN can yield better generalization scores with fewer input indices. However, the difference in scores was only between 0.03 and 0.05, showing that DFA was also able to produce good generalization on the data set. This small difference may suggest that interactions between different indices are not very important for determining the type of sequence in these models. The importance of

interactions probably depends on the type of data encoding used: a model that considers entire motifs, for instance, might benefit from inclusion of interactions between the identity of bases at different positions (see Chapter 2). The (mainly compositional) indices used here may not influence each other in the same way, thus diminishing the importance of interactions.

The choice of indices used in this experiment was not determined by any *a priori* experimentally derived knowledge about the regulatory systems in *E. coli* and *S. solfataricus*. Instead, a complete enumeration of very basic sequence and structural features was undertaken to simulate a naïve approach that could be applied to a poorly studied genome. However, if some information on regulatory binding sites within a given genome were available, as they are for the genomes considered here, then the initial indices could be extended or refined to reflect this knowledge. For instance, the frequency and position of specific degenerate oligonucleotide patterns of size > 3 nt could be determined and converted to indices. Other indices could represent weight matrix or hidden Markov model scores, if well characterized patterns were targeted. This approach would allow the application of methods like phylogenetic footprinting and comparative genomics (Gelfand *et al.*, 2000; Mironov *et al.*, 1999) to the OGA/GANN system.

Within the context of the naïve approach, other indices could also be considered to improve the prediction accuracy. The helical transformation parameters used here are derived from Gorin *et al.* (1995) and are based on analysis of many crystal DNA structures. The crystallization experiments that generated these data often used different experimental solutions and conditions, which are likely to impact the results (Dickerson *et al.*, 1994). However, structural indices could also be generated from theoretical models or other types of empirical data such as NMR (Gavathiotis *et al.*, 2000).

[5]

The Motif Genetic Algorithm as a Tool for Detecting Conserved Patterns in Biological Sequences

Motivation

The experimental approach described in Chapter 4 relied on indices of predetermined size and position, with each index representing a single sequence or structural characteristic of DNA. While some of the indices thus derived could be associated with known biological patterns, the initial generation of these indices was not based on examination of patterns within the sequences to be classified. This approach is limited in two important ways: first, the window sizes and overlap represent informed guesses as to the variability in positioning of conserved sequences; and second, any window that contains all or part of a real motif must represent the motif by its sequence and structural composition, and cannot represent the motif directly as a discrete set of ordered sequence and structural patterns. From a computational standpoint, the complete enumeration of simple characters yields a large set of indices, many of which are useless in sequence classification and prediction. The task of the OGA is thus confounded by the need to search through many patterns that make no contribution to prediction accuracy. Complete enumeration of indices also prohibits the explicit consideration of longer oligonucleotides and combinations of parameters in a single index.

As described in the Introduction, many statistical methods have been used to identify conserved patterns in biological sequences. While these methods can detect biologically relevant motifs, they are constrained by their inability to combine different types of sequence and structural information into a single motif representation. Typically, the methods focus on the order of residues, with some allowance for degenerate characters or categories, but do not consider more than one type of information at a time.

The motif genetic algorithm (MGA) is a procedure that attempts to remedy the biological and computational limitations of a completely enumerated data set. The purpose of the MGA is to pre-screen the DNA sequences and identify specific motifs that are more abundant in one set of sequences than in another. The key characteristic of the MGA design is that motifs are represented as a series of distinct characters. The set of these characters can be used as a search key in a sequence, but the individual characters, and the evolutionary operations that can be applied to them, give the MGA its flexibility in the development and detection of conserved motifs.

Design and Implementation

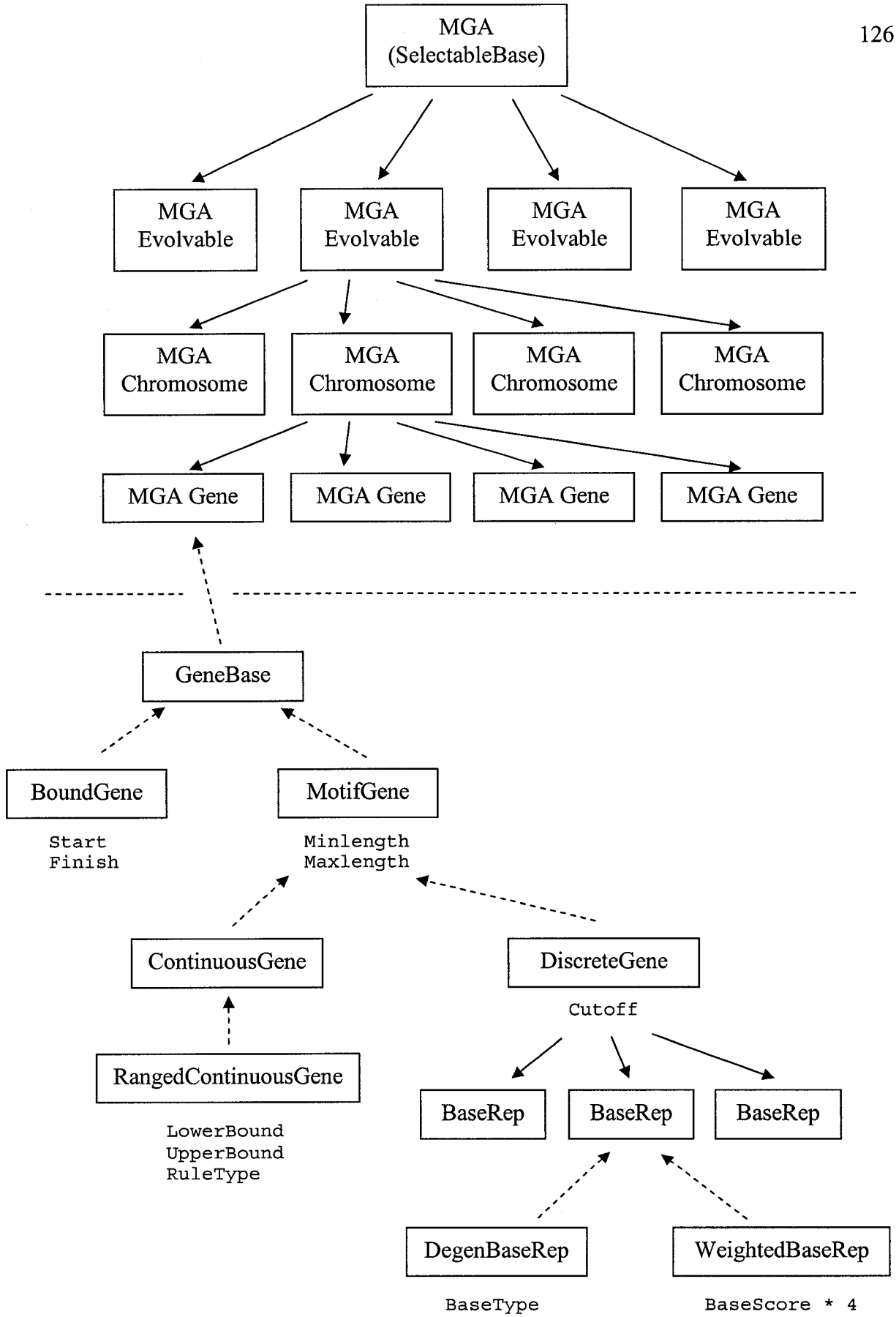
Complex Motifs: Construction and Searching

Like the OGA and IGA described in Chapter 2, the MGA is based on the C++ genetic algorithm classes initially developed by Robert L. Charlebois. The MGA architecture is summarized in Figure 5-1. A given MGA is an object of type SelectableBase which represents a set of interbreeding populations (the MGA Evolvables), with each separate population comprised of MGA Chromosomes. Each MGA Chromosome is interpreted as a motif which can be searched for in a target set of sequences. Finally, each MGA Chromosome contains a set of MGA Genes, which are described in the most general sense as objects of type 'GeneBase'.

Each GeneBase object within a given MGA Chromosome represents a single component of the specified motif. A single component can be any of the following types, which are all derived from GeneBase:

- ContinuousGene – specifies the minimum and maximum length of a 'gap' in the motif.
- RangedContinuousGene – specifies a series of bases that have a value within a specified range for a specific property such as helical twist or deformation energy. A RangedContinuousGene also specifies a minimum and maximum number of bases that are required to have this property.

Figure 5-1: Class representation of the Motif Genetic Algorithm (MGA). The boxes above the dashed line indicate the organization of populations of motifs, where an MGA object will contain many Evolvables, each containing an interbreeding population of Chromosomes. Each Chromosome contains a number of genes, which represent components of the motif specified by the Chromosome. Below the dashed line, dashed arrows show the specialization of different GeneBase objects. BoundGene is a special type of GeneBase object that represents the start and stop points for the motif search in a DNA sequence. MotifGenes specify motif components, where a ContinuousGene is a 'gap' with a specified length that can have any sequence and structural properties, while a RangedContinuousGene imposes a requirement for the value of a given property to fall within a specified range. DiscreteGenes describe a series of zero or more bases, each represented by an object of type BaseRep. Finally, each BaseRep object will be either of type DegenBaseRep, where an exact character match is required for a score of 1.0, or of type WeightedBaseRep, where each base has an associated score between 0.0 and 1.0.


MGA
(SelectableBase)
MGA
Evolvable
MGA
Evolvable
MGA
Evolvable
MGA
Evolvable
MGA
Chromosome
MGA
Chromosome
MGA
Chromosome
MGA
Chromosome
MGA Gene
MGA Gene
MGA Gene
MGA Gene
GeneBase
BoundGene
Start
Finish
MotifGene
Minlength
Maxlength
ContinuousGene
DiscreteGene
Cutoff
RangedContinuousGene
LowerBound
UpperBound
RuleType
BaseRep
BaseRep
BaseRep
DegenBaseRep
BaseType
WeightedBaseRep
BaseScore * 4

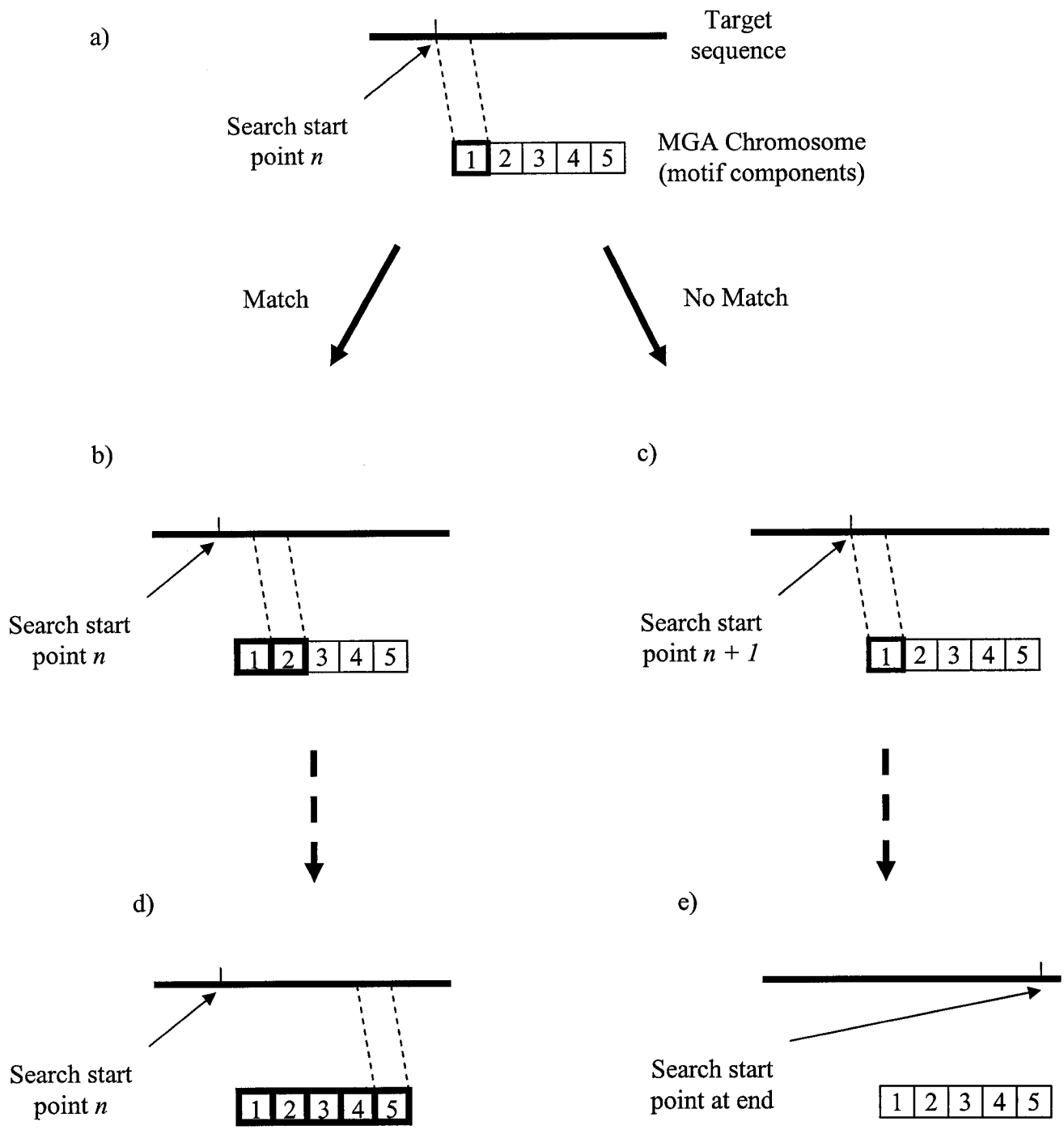
- DiscreteGene – A sequence of zero or more bases, each described in one of the two following ways:
    - DegenBase – A nucleotide base represented in IUPAC form, can be specific (A,C,G,T) or degenerate (S,W,R,Y,K,M,B,D,H,V,N)
    - WeightedDegenBase – A nucleotide base with four associated scores, one for each nondegenerate base type.

  The DiscreteGene object has a cutoff value between zero and the length of the specified sequence. A given target sequence will be assigned a summative score, with 1.0 added for each perfect match to a DegenBase, and the appropriate base score added for each WeightedDegenBase. The score thus obtained must be greater than the cutoff value for the target sequence to match with the DiscreteGene object.
- BoundGene – A special type of gene that specifies the first and last positions to start searching for the specified motif in a given DNA sequence. Each MGA Chromosome has exactly one BoundGene.

The first three GeneBase types listed above are collectively referred to as MotifGenes, since they all specify motif components.

An MGA Chromosome-specified motif is searched against a target sequence by comparing each GeneBase object to the appropriate bases in the sequence (Figure 5-2). For each possible starting point in the target sequence (likely restricted by the BoundGene), a recursive function is used to search through each motif component in turn, and to consider each possible length of ContinuousGene and RangedContinuousGene objects between the maximum and minimum. Each GeneBase object is compared to the sequence in turn, and the next GeneBase object in the MGA Chromosome is only searched if the current one matches the sequence. If the final motif component is reached and also matches the target sequence, then a match for the entire motif is recorded. When a motif match is found, the next search against the target sequence will start at the position after the end of the previous motif match. If a match is

Figure 5-2: Search strategy for MGA. The search begins at a starting point within the sequence that is determined by the BoundGene. *a)* The first motif component is compared to one or more nucleotides in the target sequence. *b)* If the target sequence satisfies the requirements of the motif component, then a match occurs and the nucleotides immediately downstream of the match are compared to the second motif component. *c)* If no match is recorded, then the search start point in the sequence in incremented by 1 and motif component 1 is compared to the new subsequence. The search continues until a match to the final motif component is recorded (*d*), or until the end of sequence is reached (*e*).


a)
Target sequence
Search start point *n*
1 2 3 4 5
MGA Chromosome (motif components)
Match
No Match
b)
Search start point *n*
1 2 3 4 5
c)
Search start point *n + 1*
1 2 3 4 5
d)
Search start point *n*
1 2 3 4 5
e)
Search start point at end
1 2 3 4 5

not found at a given sequence position, then the search will start again from the next base in the target sequence.

Fitness

The fitness of a given MGA Chromosome is determined by comparing the frequency of matches to a positive set and a negative set. A single value x_a is calculated for the negative set, which represents the average number of matches in each negative set sequence. A similar calculation is performed on the sequences in the positive set, but instead of considering the set as a whole, the mean score $x_{b(i)}$ is calculated for a small number of sequences, randomly sampled from the positive set. This random sampling is carried out a number (n) of times, generating a group of mean counts $x_{b(0)}, x_{b(1)}, \ldots, x_{b(n)}$ for the positive set. A subset of size k from this group, containing the k largest values of $x_{b(i)}$, is then averaged to yield a single positive set score, x_p. The fitness w of the MGA Chromosome is then expressed as:

$$w = [(x_p - x_a) / (x_a + c)](S(K) + 1.0) \qquad (5.1)$$

which will yield larger values when $x_p >> x_a$, and when x_a is small. The 'c' parameter is simply a constant value added to ensure that the denominator of the equation is never 0. If the value of c is very small, then small changes in x_a will have a larger impact on overall fitness, and motifs that yield no false positives will be strongly favoured. If c is large, the role of the denominator will be decreased, and false positives will be tolerated if the MGA model detects a high proportion of true positives as well.

Subsampling is meant to be used in situations where the patterns are not expected in all positive set cases. Rare patterns will be overrepresented in some subsamples, thus creating the possibility of detecting them above background noise. If a pattern is thought to be in most or all members of the positive set, then the whole positive set can be presented as a single sample.

The value S in the fitness equation represents the *specificity* of the motif. This quantity is determined by the length and degree of discrimination associated with a given

MGA Chromosome. The specificity of an MGA Chromosome is determined by summing the specificity of its component Genes. Each type of gene has its own associated specificity function:

- ContinuousGenes represent gaps in the sequence, and therefore contribute nothing to specificity.
- RangedContinuousGenes specify a physical property with a given length and a constrained range. To calculate specificity, the allowed range of the property is divided by the entire range, and this value is subtracted from 1.0 and multiplied by the minimum length of the feature.
- DiscreteGenes represent nucleotide bases either in IUPAC format or in a weight matrix representation as described above. The specificity of an IUPAC base is 1.0 for nondegenerate bases, 2/3 for twofold degenerate bases, and 1/3 for threefold degenerate bases. For a weight matrix representation, the sum and the sum of squares are calculated for the four base scores; the specificity is equal to the sum of squares divided by the sum. The overall specificity of a DiscreteGene can be no greater than the cutoff value minus the number of automatic matches (N bases).

For all Gene classes, the maximum contribution that any single position can make to overall specificity is 1.0. The K parameter is a constant value that determines the impact of specificity on fitness, and is set prior to the start of an MGA run.

Evolutionary Operations

As with the OGA and IGA, the MGA Chromosomes are subject to evolutionary operations based on their relative fitness. Parameters whose values can be set by the user are indicated in capitals.

MOVESTART – While the GeneBase objects in an MGA Chromosome are represented as an ordered series of motif components, the first component searched will not necessarily be the first in the series. The Chromosome contains a variable that specifies which component to consider first when searching for an entire motif. If the first

GeneBase to consider is not the first in the series, then the motif search will wrap around to the first in the series when it reaches the last GeneBase. For instance, if three GeneBase objects gb_1, gb_2, and gb_3 specify the bases A, G, and T, respectively, then moving the search start to position 2 will start the motif search at gb_2 and search for the motif 'GTA'. A fraction of all Chromosomes in an Evolvable are considered for this modification after each training round, and each of these Chromosomes will have an associated probability of undergoing a change in the search start from the current GeneBase to one of its neighbours. This function allows MGA to introduce new features upstream or downstream of important motif components by importing them from the other end of the Chromosome.

RECOMBINE – After the set of MGA Chromosomes within a single Evolvable have been sorted by their fitnesses, randomly selected pairs of Chromosomes from the best-scoring fraction will each contribute some of their GeneBase components to recombinant 'offspring'. These offspring will then replace the worst scoring fraction of Chromosomes in the population. If the best and worst fractions overlap, then some of the parents that contribute to the next generation will be replaced by recombinants. A single MGA Chromosome can serve as both 'parents' of a recombinant, which will yield an offspring Chromosome identical to the parent. This clonal offspring may then be subject to mutation (see below).

MUTATE (bounds) – Each MGA Chromosome contains a unique BoundGene, which specifies the legal range of positions within a target sequence that can serve as start points for the motif search. After recombination, a given fraction of the Chromosome population will be subject to BoundGene mutations. Within each Chromosome in the specified range, there is a probability that either the lower or upper bound of the legal range will be modified by up to a specified amount.

MUTATE (type) – A given fraction of MGA Chromosomes will be subject to 'type' mutations. For all of the MotifGenes within each Chromosome in this fraction, there is a probability that the MotifGene will be deleted and replaced with a new MotifGene of a random type. These substitutions typically preserve the length of the motif component while changing the required sequence or structural characteristic. However, if the length of a mutated motif component is 0, then a new BaseGene of any

length may be introduced. Length conservation is an optional feature, and may be inactivated.

MUTATE (parameter value) - A given fraction of MGA Chromosomes will be subject to parameter value mutations. For all of the MotifGenes within each Chromosome in this fraction, there is a probability that a property of the motif component will be modified. Properties that can be modified are:

- Component length for all MotifGenes;
- The upper and lower bounds of the acceptable range of values in a RangedContinuousGene;
- The base type required at a specific position within a DiscreteGene. If a DegenBase is mutated, the required character can undergo either a 'horizontal' mutation which preserves the degree of degeneracy (i.e., from A to T or B to D), or a vertical mutation which yields either a more specific or less specific base type (i.e., from A to R or B to Y). If a WeightedDegenBase is mutated, then the score associated with one of the four bases will increase or decrease.
- The cutoff value associated with a DiscreteGene.

A separate parameter determines the magnitude of the parameter value mutation in all instances above.

JOIN and SPLIT – A given fraction of MGA Chromosomes will be subject to joining or splitting of any DiscreteGenes they contain. With a certain probability, a GeneBase object (g_1) and its neighbour (g_2) will be considered for joining or splitting. If g_1 and g_2 are both DiscreteGenes, then the set of bases specified by g_2 will be added to the end of g_1's bases, and g_2 will be deleted and replaced with a spacer of length 0 to preserve the overall length of the MGA Chromosome motif. If g_1 is a DiscreteGene and g_2 is any MotifGene with a length of 0, then some of the bases may be removed from the end of g_1 and used to create a new DiscreteGene that replaces g_2. This splitting and substitution also preserves the length of the MGA Chromosome motif. These operations allow a series of conserved bases to be united within a single motif component, thus freeing up another component for further recombination and mutation.

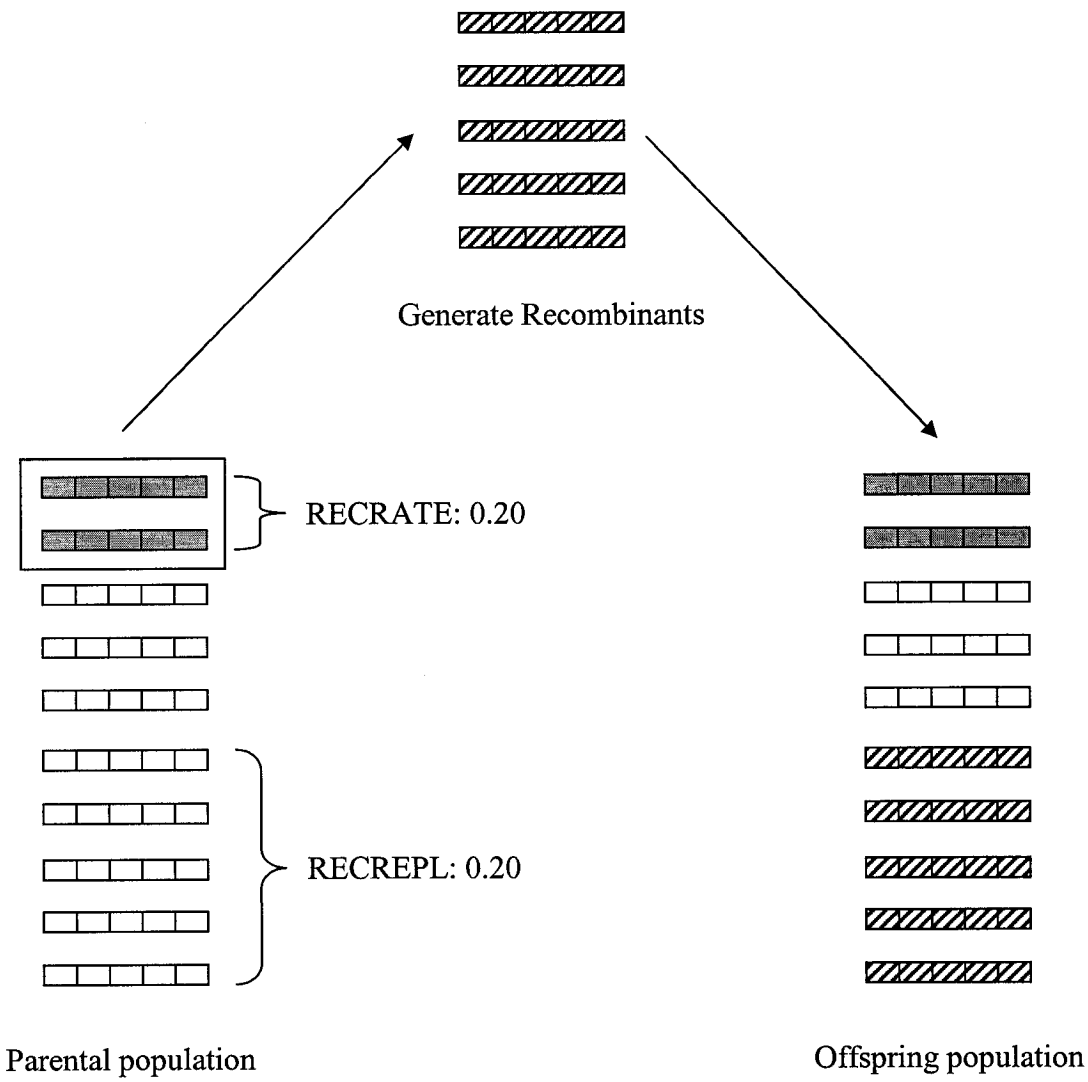
SWAP – A given fraction of MGA Chromosomes will be subject to reciprocal exchange (swapping) of two of their adjacent MotifGene objects. For each adjacent pair of MotifGenes, there is a specified probability that they will be exchanged with one another if the length of one or both MotifGenes is 0.

MIGRATE – A proportion of the MGA Chromosomes in one Evolvable can be exported to another Evolvable, where they will be able to recombine with the members of another interbreeding population. While replacement of parental MGA Chromosomes with recombinants reduces genetic diversity within an Evolvable, migration and mutation can increase it and provide new combinations of search criteria.

CRUSH – The program can introduce a bias that favours replacement of MotifGenes of any type with spacers (ContinuousGenes) of the appropriate length. The MotifGenes from a fraction of MGA Chromosomes will be subject to this replacement, with a predetermined probability. The goal of this operation is to eliminate MotifGenes that contribute needless complexity to the motif specified by an MGA Chromosome.

The 'Prop' variables described in many of the operations above, which modify existing Chromosomes through methods other than recombination, are always applied to the 'lower' end of the sorted list of MGA Chromosomes. This fraction is most likely to be comprised of new recombinants that have recently replaced low-fitness Chromosomes, and possibly some low-fitness Chromosomes that were not replaced. The preferential mutation of these MGA Chromosomes introduces more diversity and new combinations without altering the highest-scoring Chromosomes. Setting any of the 'Prop' variables to a value of 1.0 would make the entire population of MGA Chromosomes subject to the relevant operation, but could have the deleterious effect of disrupting good motifs in the highest-scoring Chromosomes. Conversely, setting any of these variables to a value of 0.0 inactivates that operation. An example is shown with recombination in Figure 5-3.

Figure 5-3: Recombination of MGA Chromosomes. Two variables determine the extent of recombination and replacement: the first (RECRATE) determines the fraction of parental MGA Chromosomes that will recombine, and the second (RECREPL) determines the percentage of parentals that are replaced by the new recombinants. The parental Chromosomes are first sorted in descending order of fitness, then enough recombinants are built from the best parentals to replace the worst parentals. Other variables that determine the fraction of a population subject to processes such as mutation are applied to the population in the same way as RECREPL.



Notation and Examples

The best-scoring MGA Chromosomes from each Evolvable are printed out after each training round. The parameters of each GeneBase component are printed out, and adjacent GeneBase objects are separated by vertical lines or pipes. Since each type of GeneBase object has a different set of information, the output format is specific to the type of gene.

BoundGene: StartPoint-EndPoint
ContinuousGene: Nminlength-maxlength
RangedContinuousGene: RULENAME Lminlength-maxlength * Vminval-maxval
DiscreteGene: baseRep/baseRep/.../cutoff
DegenBaseRep objects are represented with the appropriate IUPAC letter, while WeightedBaseRep objects are represented in the form Aval Cval Gval Tval.

For instance, the motif 'TATAAT' could be represented as follows:

```
185-234|N1-1|T/A/2|T/A/A/T/3.514|N1-1|N0-0|N0-0|N0-0|N0-0
```

The first two numbers indicate the range of start points to consider when searching for the motif in a target sequence. A gap of size 1 is followed by a 'TA' dinucleotide, which must match perfectly. The next component is a 'TAAT' tetranucleotide, which must also match perfectly since DegenBaseReps can only match with a score of 0.0 or 1.0. Since the cutoff is greater than 3.0, all four bases must match. The rest of the motif is merely a spacer of length 1.

A more complex motif could be represented in the following way:

```
72-122|N0-0|T/1|N0-0|N0-0|X(0)|A0.3975 C0.4095 G0.8614
T0.1282/A/1.761|SOMERULE L2-3 * V2.0-6.5|N1-1
```

In this case, the motif search will start at positions 72 to 122 in the target sequence. A perfect match to 'T' is required, and the surrounding spacers will match automatically.

The notation 'X(0)' represents a DiscreteGene of length zero, which therefore encodes no bases and also matches automatically. The sequence must then match the dinucleotide 'GA', since G is the only nucleotide with a score greater than 0.761 (with the following 'A' contributing 1.0 to the score of this component). After the GA match, the next 2 or 3 bases must have an average SOMERULE property between 2.0 and 6.5, where SOMERULE is a structural conversion from oligonucleotides to floating-point values. Finally, the match to the final gap of length 1 is automatic.

While the choice of output format is arbitrary, the format described above is a clear and accurate way of describing the motifs encoded by an MGA Chromosome. It can be condensed without loss of motif information by removing any components of length 0, and by consolidating adjacent gaps. Finally, if the length specified by a RangedContinuousGene is shorter than the minimum sequence length required to convert a sequence into a numerical representation, then that RangedContinuousGene can be represented as a gap. For instance, if the SOMERULE conversion was based on a set of tetranucleotide rules, then the motif component above is meaningless since it cannot be converted. Therefore, a condensed version of the second motif would be:

```
72-122|T/1|A0.3975 C0.4095 G0.8614 T0.1282/A/1.761|N2-3|N1-1
```

and since gaps at the far end are meaningless, the motif could be further simplified to:

```
72-122|T/1|A0.3975 C0.4095 G0.8614 T0.1282/A/1.761
```

Figure 5-4 shows the format that will be used to display motifs in the remainder of Chapter 5 and in Chapter 6. This format allows a quick horizontal read of the motif, and organizes the characteristics of complex components such as weight matrix representations and structural Genes.

Figure 5-4: Visual representation of an MGA Chromosome. For sequence-based motif components, the string of consecutive IUPAC (*a*) and/or weight matrix positions (*b*) are shown in order, with the minimum score required for a match shown below the motif component. IUPAC letters and their corresponding bases are shown in Table 1-1; a base that matches the appropriate IUPAC letter is assigned a score of 1.0, while a mismatch is assigned a score of 0.0. A weight matrix position shows the score associated with adenine, cytosine, guanine, and thymine in four rows. (*c*) Gaps are shown with their associated length (L:) in nucleotides. (*d*) Structural motif components also have an associated length, as well as an acceptable range (R:) for the averaged parameter value (See appendix A for structural mapping rules). The bottom row of a structural component indicates what percentage of all possible *n*-mers of length L are within the acceptable range R.

a)

A	D
2.0	

b)

1.0	0.2
0.2	0.5
0.1	0.9
0.5	0.0
1.4	

c)

GAP
L: 2-4

d)

TWIST
L: 3
R: 32.3/35.4
25%

Tests with Artificial Sequences

A series of experiments were carried out to test the ability of MGA to handle different types of pattern detection problems. Separate MGA runs were performed to address four key properties of regulatory sequences: variable positioning within upstream regions, imperfect adherence of individual sequences to the consensus pattern, motifs present in only a subset of all upstream regions, and the importance of DNA structure. While the tests described below simplify the biological problem, they are vital in assessing the ability of MGA to detect patterns, and in showing how MGA parameters can be tailored to address specific problems.

For all of the runs described below, a positive and negative set of DNA sequences were generated, each with 100 members. Each of these sequences was 200 nt in length, and each of the four nucleotide bases had an equal probability of occurring at each position within the sequences. Predefined motifs were then inserted into the positive set, with the type and positioning of each motif specific to the type of run to be performed.

Most MGA parameters were unchanged over the entire set of experiments. Table 5-1 shows a complete list of evolutionary parameters, with the values used for each of the four runs. Where parameters differ, the justification for the change will be given in the run descriptions below. Each MGA experiment was performed for 100 generations, with 100 MGA Chromosomes per Evolvable. The MGA Chromosomes were comprised of 8 motif components. An important parameter is *c,* the constant in the fitness function denominator described in *Fitness* above. Since the sequences used in the experiments below all contain the relevant inserted patterns, and the patterns are reasonably strong, *c* was set to a very low value (0.01) to strongly favour models that yielded no false positives. To simplify the generated models, DiscreteBase objects contained only IUPAC base representations in these experiments.

Experiment #1 – Variable Positioning of a Single Motif

In this experiment, each sequence in the positive set was 'spiked' with a completely conserved motif. The *E. coli* σ^{70} consensus (TTGACA(N_{17})TATAAT) was

Table 5-1: Summary of parameter values used in four MGA experiments performed to detect artificial motifs. Detailed descriptions of the relevant MGA functions can be found in the **Design and Implementation** section of this chapter. Abbreviated names of important parameters are shown in the second column, and parameters whose values were different in one or all of the experiments are indicated with boldface text.

Parameter Description	Parameter Name	Value (Exp. 1)	Value (Exp. 2)	Value (Exp. 3)	Value (Exp. 4)
Number of MGA Evolvables	**NumEvo**	**50**	**50**	**100**	**50**
Size of each positive set sample	**SampleSize**	**100**	**100**	**10**	**100**
Number of random samples	**SampleNum**	**1**	**1**	**100**	**1**
Fraction of samples that contribute to fitness	**SampFrac**	**1.0**	**1.0**	**0.1**	**1.0**
Starting frequency of different motif component classes	**Gap**	**60%**	**60%**	**60%**	**40%**
	IUPAC	**40%**	**40%**	**40%**	**30%**
	RangedCG	**0%**	**0%**	**0%**	**30%**
Starting length of PosGenes	StartLen	75	75	75	75
Maximum DiscreteGene length	MaxDisc	6	6	6	6
Maximum Gap length	MaxGap	20	20	20	20
Maximum RangedContinuousGene length	**MaxRCG**	**N/A**	**N/A**	**N/A**	**12**
Fraction of parental Chromosomes that recombine	RecRate	0.2	0.2	0.2	0.2
Fraction of Chromosomes replaced by recombinants	RecRepl	0.6	0.6	0.6	0.6
BoundGene mutation rate	MutRateB	0.4	0.4	0.4	0.4
BoundGene maximum mutation amount	MutAmtB	100 nt	100 nt	100 nt	100 nt
Fraction of MGA Chromosomes subject to BoundGene mutation	MutPropB	0.6	0.6	0.6	0.6
Rate of mutation affecting Gene type	**MutRateT**	**0.2**	**0.2**	**0.2**	**0.4**
Fraction of MGA Chromosomes subject to type mutation	MutPropT	0.6	0.6	0.6	0.6

Table 5-1, continued

Parameter Description	Parameter Name	Value (Exp. 1)	Value (Exp. 2)	Value (Exp. 3)	Value (Exp. 4)
Rate of type-conservative mutations	MutRateP	0.2	0.2	0.2	0.2
Maximum magnitude of type-conservative mutation	MutAmtP	0.2	0.2	0.2	0.2
Fraction of MGA Chromosomes subject to type-conservative mutation	MutPropP	0.6	0.6	0.6	0.6
Probability of changing the first Gene tested by the MGA Chromosome	MoveRate	0.2	0.2	0.2	0.2
Fraction of MGA Chromosomes subject to change in first Gene	MoveProp	0.4	0.4	0.4	0.4
Probability of joining or splitting adjacent motif components	JoinRate	0.6	0.6	0.6	0.6
Fraction of MGA Chromosomes subject to joining or splitting	JoinProp	0.2	0.2	0.2	0.2
Probability of swapping adjacent MGA Chromosome Genes	SwapRate	0.3	0.3	0.3	0.3
Fraction of MGA Chromosomes subject to Gene swapping	SwapProp	0.6	0.6	0.6	0.6
Rate of MGA Chromosome exchange between Evolvables	**MigRate**	**0.1**	**0.1**	**0.0**	0.1
Probability of converting a MotifGene to a gap	CrushRate	0.1	0.1	0.1	0.1
Fraction of MGA Chromosomes subject to gap replacement	CrushProp	0.6	0.6	0.6	0.6
Value of constant in the fitness function denominator	*c*	0.01	0.01	0.01	0.01
Contribution of specificity to fitness	**K**	**0.2**	**0.2**	**0.2**	**0.001**

inserted into each sequence exactly once, starting at a randomly selected position between 50 and 100 nt from the beginning of the 200 nt sequence.

Since a single motif was inserted into every positive set sequence, a single representation is sufficient to account for all positive set members. The migration rate was set to 0.1 to allow passing of MGA Chromosomes between different Evolvables, yielding a more homogeneous population. The homogeneity of the positive set also eliminates the need for subsampling, so the fitness was calculated from a single sample that contained the entire positive set.

Figure 5-5 shows the fittest MGA Chromosomes from the best and worst Evolvables after 100 generations. The best Chromosome from the worst Evolvable has only 66% of the fitness of the best Chromosome overall. Despite this difference, even the less fit Chromosome clearly detected part of the conserved motif. The fitness of this Chromosome appears to be limited by the restricted range imposed by the PosGene. While the 'TATAAT' motif recognized by this Chromosome could be present anywhere between sequence positions 73 and 123 in the positive set, the PosGene of this Chromosome limited the start point of the search between positions 97 and 124. The best Chromosome detected the entire 12 nt motif, and was able to accurately model the spacer region between the two components. The starting positions considered by the PosGene cover all of the possible motif positions. This motif is not optimal, because there are two unnecessary degenerate bases within the 12 conserved bases. However, the Chromosome model was sufficiently specific to yield perfect discrimination between the positive and negative sets.

Experiment #2 – Degenerate Variants of a Motif

To address the issue of degeneracy in motifs, modified versions of the σ^{70} consensus were introduced into another set of randomly generated sequences to yield a new positive set. The variants of both halves of the motif are shown in Figure 5-6*a*; each inserted motif combined a randomly selected TTGACA variant with a randomly selected TATAAT variant. The starting position of each motif was fixed at base 75 in the 200 nt sequences to eliminate the confounding effect of variable positioning.

Figure 5-5: The inserted consensus sequence (*a*) and the fittest MGA Chromosomes from the Evolvables with the lowest (*b*) and highest (*c*) fitnesses in Experiment #1. For sequence-based motif components, the string of IUPAC and/or weight matrix positions are shown in order, with the minimum score required for a match shown below the motif component. IUPAC letters and their corresponding bases are shown in Table 1-1; a base that matches the appropriate IUPAC letter is assigned a score of 1.0, while a mismatch is assigned a score of 0.0. A weight matrix position shows the score associated with adenine, cytosine, guanine, and thymine in four rows. Gaps are shown with their associated length (L:) in nucleotides. Structural motif components also have an associated length, as well as an acceptable range (R:) for the averaged parameter value (See appendix A for structural mapping rules). The bottom row of a structural component indicates what percentage of all possible *n*-mers of length L are within the acceptable range R. In Figures 5-5*b* and 5-5*c*, the sites that are consistent with the inserted consensus are underlined, and the positions they match (if any) are indicated beneath the appropriate motif component.

a) Consensus:

Position	1	2	3	4	5	6	7-23	24	25	26	27	28	29
Base	**T**	**T**	**G**	**A**	**C**	**A**	**N**	**T**	**A**	**T**	**A**	**A**	**T**

b) Worst MGA Chromosome:

T	A	T	A	A	T
6.0					
24	**25**	**26**	**27**	**28**	**29**

c) Best MGA Chromosome:

T	T	G	A	C	R	**GAP**	T	A	T	R	A	W
						L: 17-18						
6.0							**6.0**					
1	**2**	**3**	**4**	**5**	**6**	**7-23**	**24**	**25**	**26**	**27**	**28**	**29**

As in Experiment #1 above, a single model should be sufficient to represent all motifs in the positive set. Migration was again enabled, and multiple subsampling was not used to assess Chromosome fitness. The starting ratio of IUPAC bases to gaps was also preserved.

Figure 5-6 shows the fittest MGA Chromosomes from the best and worst Evolvables after 100 generations. While 49 out of 50 Evolvables had fitnesses between 2.00 and 2.53, the best MGA Chromosome from the worst Evolvable had a fitness of only 1.44. The main component of this MGA Chromosome is the requirement to match 4 out of 5 bases of the sequence VTWTA. This sequence matches the first four bases of Motif Component 2, which are composed almost entirely of the Weak bases A and T. The best MGA motif clearly matches components from both halves of the degenerate motif. This latter MGA Chromosome recognized all 100 members of the positive set, while matching 4 out of 100 negative set members. The best MGA Chromosomes exploit the flexibility of degenerate IUPAC characters and the requirement for only partial sequence matches (implemented *via* the threshold) to model all of the motif variants.

Experiment #3 – Motifs in a Subset of All Positive Sequences

The positive sequence set for the third experiment was generated by introducing one of six different motifs (Figure 5-7*a*) into each 200 nt sequence. Each of these six motifs was completely conserved, and there was no intermixing of components. As in Experiment #2, the motifs were always introduced at position 75 of the positive set sequences.

These motifs have no clear relationship, so a single model should not be used to represent all of them. Since homogenization of the population is not desired in this case, the migration rate in this experiment was set to 0.0. The number of Evolvables was increased to 100, to increase the likelihood of all six motifs being represented in the final set. Finally, since each motif was present in only 16.6% of positive set sequences on average, subsampling was used to overrepresent these signals. Fifty samples of 10

Figure 5-6: The inserted sequences (*a*) and the fittest MGA Chromosomes from the Evolvables with the lowest (*b*) and highest (*c*) fitnesses in Experiment #2. For sequence-based motif components, the string of IUPAC and/or weight matrix positions are shown in order, with the minimum score required for a match shown below the motif component. IUPAC letters and their corresponding bases are shown in Table 1-1; a base that matches the appropriate IUPAC letter is assigned a score of 1.0, while a mismatch is assigned a score of 0.0. A weight matrix position shows the score associated with adenine, cytosine, guanine, and thymine in four rows. Gaps are shown with their associated length (L:) in nucleotides. Structural motif components also have an associated length, as well as an acceptable range (R:) for the averaged parameter value (See appendix A for structural mapping rules). The bottom row of a structural component indicates what percentage of all possible *n*-mers of length L are within the acceptable range R. In Figures 5-6*b* and 5-6*c*, the sites that are consistent with the inserted consensus are underlined, and the positions they match (if any) are indicated beneath the appropriate motif component.

a) Inserted sequences:

Position	1	2	3	4	5	6	7-23	24	25	26	27	28	29
Motif 1	**T**	**T**	**G**	**A**	**C**	**A**	**N**	**T**	**A**	**T**	**A**	**A**	**T**
Motif 2	**T**	**A**	**G**	**A**	**C**	**A**	**N**	**A**	**A**	**T**	**A**	**A**	**T**
Motif 3	**T**	**T**	**T**	**A**	**C**	**A**	**N**	**T**	**A**	**C**	**A**	**A**	**T**
Motif 4	**T**	**T**	**G**	**A**	**G**	**A**	**N**	**T**	**A**	**C**	**A**	**G**	**A**
Motif 5	**T**	**A**	**T**	**A**	**G**	**A**	**N**	**T**	**A**	**T**	**A**	**C**	**C**
Motif 6	**T**	**T**	**G**	**A**	**G**	**T**	**N**	**A**	**A**	**T**	**A**	**A**	**G**
Motif 7	**T**	**T**	**G**	**A**	**C**	**T**	**N**	**T**	**A**	**C**	**A**	**A**	**C**
Motif 8	**T**	**T**	**G**	**A**	**C**	**G**	**N**	**T**	**A**	**T**	**A**	**G**	**G**
Motif 9	**T**	**T**	**T**	**A**	**G**	**T**	**N**	**T**	**A**	**T**	**A**	**C**	**T**
Motif 10	**T**	**T**	**T**	**A**	**G**	**A**	**N**	**A**	**A**	**T**	**A**	**C**	**T**
Summary	**T**	**W**	**K**	**A**	**S**	**D**	**N**	**W**	**A**	**Y**	**A**	**V**	**N**

b) Worst MGA Chromosome:

V	T	W	T	A
		4.0		
23	**24**	**25**	**26**	**27**

c) Best MGA Chromosome:

T	T	K	A	S	W	**GAP**	V	T	W	W	A
						L: 14-16					
5.0							**4.0**				
1	**2**	**3**	**4**	**5**	**6**	**7-23**	**23**	**24**	**25**	**26**	**27**

Figure 5-7: The inserted sequences (*a*) and the best-scoring MGA motif for each of the six positive set motifs (*b*-*g*) in Experiment #3. For sequence-based motif components, the string of IUPAC and/or weight matrix positions are shown in order, with the minimum score required for a match shown below the motif component. IUPAC letters and their corresponding bases are shown in Table 1-1; a base that matches the appropriate IUPAC letter is assigned a score of 1.0, while a mismatch is assigned a score of 0.0. A weight matrix position shows the score associated with adenine, cytosine, guanine, and thymine in four rows. Gaps are shown with their associated length (L:) in nucleotides. Structural motif components also have an associated length, as well as an acceptable range (R:) for the averaged parameter value (See appendix A for structural mapping rules). The bottom row of a structural component indicates what percentage of all possible *n*-mers of length L are within the acceptable range R. In Figures 5-7*b* to 5-7*g*, the sites that are consistent with the inserted sequence are underlined, and the positions they match (if any) are indicated beneath the appropriate motif component.

a) Inserted Sequences:

Position	1	2	3	4	5	6	7-23	24	25	26	27	28	29
Motif 1	**T**	**T**	**G**	**A**	**C**	**A**	**N**	**T**	**A**	**T**	**A**	**A**	**T**
Motif 2	**A**	**C**	**T**	**G**	**C**	**A**	**N**	**T**	**T**	**A**	**G**	**A**	**G**
Motif 3	**G**	**G**	**G**	**C**	**A**	**A**	**N**	**A**	**A**	**C**	**G**	**T**	**A**
Motif 4	**A**	**G**	**C**	**T**	**T**	**T**	**N**	**G**	**G**	**T**	**T**	**T**	**A**
Motif 5	**C**	**C**	**C**	**C**	**C**	**C**	**N**	**G**	**A**	**T**	**C**	**G**	**A**
Motif 6	**A**	**A**	**T**	**T**	**G**	**G**	**N**	**A**	**A**	**A**	**G**	**A**	**A**

b) Motif 1:

Y	T	G	A	V	A	GAP	M	T	GAP	A	A	D	D
						L: 8-16			**L: 1-2**				
6.0							**2.0**			**3.0**			
1	**2**	**3**	**4**	**5**	**6**	**7-22**	**23**	**24**	**25-26**	**27**	**28**	**29**	**30**

c) Motif 2:

T	D	A	D	M	G	GAP	K	W	D	H
						L: 13-21				
6.0							**3.0**			
24	**25**	**26**	**27**	**28**	**29**	**30-?**				

Figure 5-7, continued

d) Motif 3:

D	W	R	R	S	GAP	M	A	C	R	W	A
					L: 11-22						
3.0				1.0		6.0					
		1	2	3	4-23	24	25	26	27	28	29

e) Motif 4:

A	V	C	T	T	K	GAP	S	G	K	K	C	M
						L: 16-18						
6.0							5.0					
1	2	3	4	5	6	7-23	24	25	26	27	28	29

f) Motif 5:

C	C	C	C	C	C	GAP	R	K	M	C
						L: 18-26				
6.0							3.0			
1	2	3	4	5	6	7-24	25	26	27	28

Figure 5-7, continued

g) Motif 6:

A	A	T	T	G	D	GAP L: 15-25	K	W	H	H	K	W	D	H
6.0							3.0				3.0			
1	2	3	4	5	6									

sequences each were tested for each MGA Chromosome, and the five (10%) best-scoring samples were used to generate the fitness score.

At the end of training, all six motifs were represented in the population of Evolvables. Figure 5-7 shows the best-scoring motif for each of the six patterns. In 5 out of 6 cases, both halves of the motif were detected and modeled, though one half was typically more specific than the other. For instance, the first half of motif #5 (CCCCCC) was modeled as such in the MGA Chromosome, but only a weak match to the second half of the motif was required. Still, this model detected all of the inserted copies of motif #5, and yielded no false positive matches. Only the second half of motif #2 (TTAGAG) was modeled by the MGA Chromosome shown here, with a very weak downstream pattern (KWDH with a matching requirement of 3 bases) that confers little to no specificity.

While exact models would have been sufficient to model the six motifs, in no case did the MGA Chromosome provide such a representation. Further training rounds might yield more specific models, but there may be a specific reason for the observed degeneracy. Since the six motifs are grouped together in the positive set, there is no selective disadvantage to an MGA Chromosome that can detect more than one of these motifs. By preserving some degree of degeneracy, the six models shown here were able to match a few sequences in the positive set that did not contain their 'primary' target motif. If the selective advantage of matching a few more motifs outweighed the reward from increased specificity, then selection would favour the evolved motifs seen here. While this tendency may be a limitation of the system, further inspection of the detected motifs following MGA analysis would quickly reveal the lack of relationship between a minority of the detected motifs. This incorrect clustering could also be addressed by changing some MGA parameters, particularly the influence of specificity on fitness and the sampling parameters.

Experiment #4 – DNA Structure Conservation

The last experiment was designed to show the potential influence of DNA structure in defining motifs. Before deciding whether to include DNA structure modeling

in MGA analysis, a critical question must be answered: can different DNA sequences yield similar structural patterns? If so, then structure can provide valuable information that is not obvious from analysis of the DNA structure alone.

To address this question, the relationship between a single sequence-dependent structural feature and the underlying DNA sequence was studied. Propellor twist describes the opposed rotation of the two bases in a pair (Dickerson, 1989), and is expressed in degrees. Baldi et al (1998) published a list of propellor twist values for each type of dinucleotide step, based on values calculated from crystallographic data (see Appendix A). One thousand random DNA hexamers were generated, and their mean propellor twist values were calculated from the list shown above. The mean values ranged from a minimum of –18.66 for AAAAAA, to a maximum of –8.11 for CCCCCC. Two blocks of ten hexamers were selected from different positions within this range, with mean propellor twist values between –11.820 and –11.808 for the first block, and between –11.328 and –11.302 for the second. The chosen hexamers are shown in Figure 5-8*a*: despite their nearly identical propellor twist values, there is little sequence similarity between the members of each set, and most sites fail to show a clear preference for any nucleotide base. The motifs inserted into the positive set sequences consisted of a hexamer from each block, separated by a 17 nt spacer. As in Experiment #3, each hexamer from the first block was associated with a hexamer from the second block.

Since a single, optimal motif was desired at the end of training, migration was enabled and subsampling of the positive set was not performed. RangedContinuousGenes were enabled to allow structural representations, but IUPAC degenerate motifs were also permitted. The MGA was thus free to choose between a loose sequence representation and a strong structural one. As a further challenge, the MGA was able to use 21 other dinucleotide structural rules in addition to propellor twist. The influence of complexity on the fitness score was reduced to 0.001 to avoid a surplus of small, only slightly restrictive structural constraints. The rate of gene type mutations was increased from 0.2 to 0.4 to increase the rate of creation of new MGA genes.

The best Chromosomes from the best and worst Evolvables are shown in Figure 5-8. Though both motifs are positioned at the correct start point, the fitness of the MGA Chromosome from the worst Evolvable is only 60.2% as high as the fitness of the best

Figure 5-8: The inserted sequences (*a*) and the fittest MGA Chromosomes from the Evolvables with the highest (*b*) and lowest (*c*) fitnesses in Experiment #4. For sequence-based motif components, the string of IUPAC and/or weight matrix positions are shown in order, with the minimum score required for a match shown below the motif component. IUPAC letters and their corresponding bases are shown in Table 1-1; a base that matches the appropriate IUPAC letter is assigned a score of 1.0, while a mismatch is assigned a score of 0.0. A weight matrix position shows the score associated with adenine, cytosine, guanine, and thymine in four rows. Gaps are shown with their associated length (L:) in nucleotides. Structural motif components also have an associated length, as well as an acceptable range (R:) for the averaged parameter value (See appendix A for structural mapping rules). The bottom row of a structural component indicates what percentage of all possible *n*-mers of length L are within the acceptable range R. In Figures 5-8*b* and 5-8*c*, the sites that are consistent with the inserted consensus are underlined, and the positions they match (if any) are indicated beneath the appropriate motif component. Where a motif component may match more than one position within the consensus, alternatives are indicated in parentheses. Propellor twist is represented by the BRUNAKPROP parameter.

a) Inserted Sequences:

	Component 1						Gap	Component 2						Mean Propellor Twist (°)	
Position	1	2	3	4	5	6	7-23	24	25	26	27	28	29	Component 1	Component 2
Motif 1	**C**	**G**	**A**	**G**	**G**	**A**	**N**	**C**	**T**	**G**	**G**	**C**	**T**	-11.82	-11.328
Motif 2	**T**	**C**	**G**	**A**	**G**	**G**	**N**	**A**	**G**	**G**	**C**	**T**	**G**	-11.82	-11.328
Motif 3	**T**	**C**	**C**	**T**	**C**	**G**	**N**	**C**	**C**	**A**	**G**	**C**	**T**	-11.82	-11.328
Motif 4	**T**	**C**	**C**	**G**	**A**	**G**	**N**	**A**	**G**	**G**	**C**	**T**	**G**	-11.82	-11.328
Motif 5	**C**	**T**	**C**	**G**	**G**	**A**	**N**	**C**	**T**	**G**	**C**	**C**	**T**	-11.82	-11.328
Motif 6	**A**	**G**	**T**	**G**	**T**	**G**	**N**	**C**	**A**	**G**	**G**	**C**	**T**	-11.82	-11.328
Motif 7	**G**	**C**	**G**	**A**	**T**	**G**	**N**	**C**	**T**	**G**	**G**	**A**	**G**	-11.81	-11.314
Motif 8	**C**	**T**	**G**	**G**	**A**	**G**	**N**	**G**	**G**	**G**	**A**	**T**	**A**	-11.808	-11.312
Motif 9	**A**	**G**	**G**	**C**	**T**	**A**	**N**	**T**	**A**	**G**	**T**	**G**	**G**	-11.808	-11.302
Motif 10	**C**	**T**	**C**	**C**	**A**	**G**	**N**	**C**	**C**	**T**	**A**	**C**	**A**	-11.808	-11.302

b) Best MGA Chromosome:

BRUNAKPROP
L: 6-7
R: -11.83/-11.73
2.77%
1-6

Figure 5-8, continued

c) Worst MGA Chromosome:

ENTROPY	BRUNAKPROP	MAJORSIZE	MAJORSIZE	GAP	V	K	W
L: 2	L: 7	L: 4-7	L: 7-9	L: 10-18			
R: -28.4/-15.2	R: -12.11/-12.00	R: 3.86/4.1	R: 3.86/4.1				
93.75%	3.58%	57.81%	93.75%		3.0		
	0-6? (1-7?)						

MGA Chromosome. The reason for this difference in fitness is apparent when the motifs are compared. The less fit motif is a collection of three different structural features (entropy, propellor twist and major groove size) and the propellor twist motif is one nucleotide too long, and the range is slightly higher than that of the first inserted motif group. This model still yields better than random predictions, but fails to identify many of the inserted motifs. In contrast, the best motif model represents the exact size and an appropriate range for the propellor twist parameter, and was able to detect all motifs in the positive set with only three false positives. The second motif component is not represented in this MGA Chromosome, and is clearly not necessary to achieve accurate predictions.

While the worst Evolvable contains a short, weak nucleotide constraint, the majority of MGA Chromosomes from the final generation had no DiscreteGenes. Since the initial probabilities of generating DiscreteGenes and RangedContinuousGenes was equal, this preference for structure strongly supports the potential role for DNA structure in motif modeling.

Conclusions

While these examples are somewhat oversimplified, they illustrate the ability of MGA to address different types of problems in motif detection, as well as the important considerations that should influence the choice of MGA parameters. Flexibility is essential in allowing MGA to handle different tasks, and the above experiments show how parameters can be changed to exploit or to anticipate certain properties of a set of motifs.

Many parameters, such as those affecting the recombination rate and the rate of joining and splitting motif components, were never modified in these analyses. If execution time were not an important consideration, then an OGA could be wrapped around this program as in Chapters 3 and 4 with the goal of determining optimal parameter values. However, optimizing MGA parameters in this way could take days or weeks, while three out of four of the above runs finished in less than an hour. An adequate compromise is to activate features such as joining, splitting, and gene crushing,

but leave the rate at low levels (0.05 to 0.2). This range allows the features to have some effect, without obscuring the benefits from the key recombination and mutation functions.

While the attempts at pattern detection were all successful and yielded very high discrimination between the positive and negative sets, the precision of the models generated *via* evolutionary means is not ideal. This effect was particularly apparent in Experiment 3, where loose models that detected a few 'other' positive set motifs were favoured, and in Experiment 4, where half of the motif was missed. An important consideration here is the influence of specificity on overall fitness. For Experiment 3, increasing the relative importance of specificity might yield fewer degenerate characters within the conserved domains. Similarly, a larger role for specificity might precipitate the addition of the second motif component in Experiment 4 by smoothing out any initial loss in fitness resulting from the creation of a new RangedContinuousGene. However, an increased role for specificity also increases the likelihood that unconserved regions will be modeled with weak constraints, since the loss of a few positive cases is offset by the gain in specificity. Such a decision may be influenced by local variation in sequence composition, rather than a biologically meaningful pattern, and yields motifs that are overly complex and inaccurate.

Pattern detection methods and statistical methods in general often rely on a null model to determine whether detected patterns are significant. This null model can be based on a statistical model of the training sequences themselves (Hertz and Stormo, 1999) or, as in MGA, on a 'negative set' of sequences thought to lack the patterns being searched for. The choice of negative set sequence is critical to the detection of relevant patterns. Since the above experiments were conducted with a background of random sequence, the only evident biases should be the deliberately inserted motifs. However, biological sequence is non-random, and there are many associated biases other than transcription factor binding sites (Pedersen *et al.*, 2000). The ideal negative set sequences will have identical nucleotide composition to the positive set sequences, but with disrupted motifs. This ideal is perhaps unachievable if the motifs are not known beforehand, as will be the case in the next chapter, but should remain the goal to yield optimal pattern detection.

[6]

Characterization of Conserved Regulatory Patterns with the Motif Genetic Algorithm

Motivation

While MGA was able to address different types of detection problems when taken separately, the detection of real biological patterns poses a greater challenge. Conserved DNA motifs exhibit more complicated combinations of the problems examined in the previous chapter, and they exist within a context of biased background sequences. Intergenic DNA is not random, and certain sequence or structural features are present with frequencies that differ greatly from expectations based only on nucleotide composition (Nussinov, 1991).

Some protein-binding sites in DNA have been extensively studied and characterized. In the prokaryotic world, the γ-proteobacterium *Escherichia coli* is currently the best-characterized organism with respect to transcriptional regulation. *E. coli* K12 has thus served as the first test for many prior detection methods (Stormo, 1990; Li *et al.*, 2002), since the predicted models can be compared with a large set of experimentally derived information. Much of this experimentally derived data is collected in the RegulonDB archive (Salgado *et al.*, 2001). RegulonDB links *E. coli* genes with the proteins that regulate them, and the upstream sites that are bound by these proteins.

In this chapter, several sets of experiments are performed to test the pattern detection ability of MGA. Models of several conserved binding sites are generated with MGA as in Chapter 5, with a general discussion of results at the end of the chapter. A key challenge in modeling regulatory protein binding sites is that the data set used for training does not reflect the entire possible set of binding sequences for a given protein. Thus, a model that recognizes those sites perfectly and excludes all others will likely miss other

valid binding sites. One method used to address this problem below exploits the heuristic nature of MGA as a possible solution to this problem: instead of training a single model that fits the training data perfectly, train a number of models independently and evaluate them all against the target sequence database. While the trained models should all focus on the same binding site and thus be similar to one another, they may rely on different signals within the same motif to make their prediction. Recognition of a target sequence by the majority of these trained models might imply stronger support for this sequence.

Constrained Models: Analysis of CAP Binding Sites

Overview

The cyclic AMP-catabolite activator protein (cAMP/CAP or simply CAP) of *E. coli* is a 'global regulator' involved in the transcriptional regulation of many genes and pathways. CAP plays a key regulatory role in the catabolism of sugars other than glucose, such as lactose, maltose, and melibiose. The genes required for transport and breakdown of these sugars are not expressed when glucose is available to the cell, but a shortage of glucose coupled with the presence of another sugar can lead to transcriptional activation of these genes. In the absence of glucose, the cAMP ligand is manufactured by the adenylate cyclase enzyme, and serves as an indicator that the cell is under nutrient stress. cAMP binds CAP and induces a conformational change that allows CAP to bind with its recognition site in DNA, where it can serve either as an activator or repressor of transcription. CAP is also involved in the regulation of genes required for functions including iron uptake, drug resistance and toxin production (Saier *et al.,* 1996). To date, CAP has been implicated in the regulation of over 150 genes in *E. coli*, through either direct transcriptional modulation of the genes themselves or through activation of other regulatory proteins (Passner and Steitz, 1997).

A number of CAP binding sites have been identified in *E. coli*, and sequence alignments reveal a consensus, 5′-TGTGA(N_{6-8})TCACA-3′. The binding site consensus

is palindromic because active CAP binds as a dimer. The strength of the interaction between CAP and a specific binding site depends on two components: the similarity of each conserved half of the motif to the consensus 5′-TGTGA-3′ (the reverse complement of which is 5′-TCACA-3′), and the length of the spacer between them, with an optimal length of 6-8 nt (Barber and Zhurkin, 1990). In addition to the pentanucleotides recognized directly by CAP, there is a preference for weak (A/T) nucleotides on either side of the binding site (Ebright *et al.*, 1989). Variation in sequence and spacer length can have a dramatic effect on the strength of a CAP site: for instance, the affinity of CAP for the consensus sequence is 500 times higher than for the site found upstream of the *lac* operon promoter (Berg and von Hippel, 1988).

Stormo and Hartzell (1989) collected a set of 18 DNA sequences, each 100 nt long and containing at least one CAP binding site. This data set was compiled from previous studies (Berg and von Hippel, 1988; de Crombrugghe *et al.*, 1984) and was used to demonstrate the effectiveness of the CONSENSUS algorithm in detecting degenerate binding sites in unaligned sequences. This data set is also used here to test the motif-finding ability of MGA. In this set of experiments, the 18 sequences containing CAP binding sites were designated as the positive set. These 18 sequences were then shuffled to disrupt the CAP binding sites while preserving the overall nucleotide composition. The negative set contained ten shuffled replicates of each positive set sequence, for a total of 180 negative set members.

Model Construction

Five separate MGA training runs were performed to detect and to model the CAP binding sites. Four of these runs were similar in that migration was enabled, thus yielding a concerted effort to find a single optimal motif, as described in Chapter 5. The fitness function used to evaluate the MGA Chromosomes in these runs was also simplified to maximize the difference between positive and negative predictions:

$$w = (x_p - x_a) \qquad (6.1)$$

where w is fitness, x_p is the mean prediction accuracy on the positive set, and x_a is the mean prediction accuracy on the negative set. The four runs were permitted different combinations of motif types, with the first constrained to IUPAC base representations, the second limited to weight matrix (WM) components, the third to structural (STRUCT) descriptions, and the fourth allowed to use all motif types (COMBINED). A fifth run was performed to study the effectiveness of a "panel of experts" model in making genomic predictions. Migration was disabled in this training run to yield a set of 50 isolated Evolvables. Since many separate Evolvables, each capable of predicting at least a few positive cases, were trained, the fitness function used was the same as in Chapter 5, with strict penalties for false positives. Five hundred rounds of MGA training were carried out in the four models with migration, and each Evolvable consisted of 100 MGA Chromosomes. Since the Evolvables in the fifth training run were isolated, the number of Chromosomes in each was increased to 200 and the number of training rounds to 1000. The parameter values used for all five training models are shown in Table 6-1.

The prediction accuracy of the best-scoring MGA Chromosomes from each of the five runs is shown in Table 6-2. Each of the four models generated with migration were able to identify either 17 or 18 of the 18 positive set sequences. However, there was a dramatic difference in the number of false positive predictions between the models created with and without structural components: the rate of false positive detection in the negative set was >10% in the best IUPAC and weight matrix models, but below 4% in the structural and combined models. This difference yielded higher overall accuracy scores for the structural models. In the experiment where the Evolvables were not subject to migration and the fitness function was designed to minimize false positives, the best overall Chromosome had a prediction accuracy of 72.2% (13/18) on the positive set, with 2.2% of all negative set members classified incorrectly. Other Evolvables had lower prediction accuracy on the training set, but better predictions on the negative set: several Evolvables correctly classified 56.6% (10/18) of positive training set members, with all negative set sequences correctly classified.

Figure 6-1 shows the best scoring model obtained from each of the five experiments. The CAP consensus site is shown in Figure 6-1*a*, and Figures 6-1*b* through 5-1*f* show the trained MGA motifs along with the consensus positions that are

Table 6-1: Summary of parameter values used in five MGA experiments aiming to detect and to model the CAP binding site. Detailed descriptions of the relevant MGA functions can be found in the **Design and Implementation** section of Chapter 5. Abbreviated parameters names are shown in the second column, and parameters whose values were different in one or all of the experiments are indicated with boldface text.

Parameter Description	Parameter Name	IUPAC Model	Weight Matrix Model	Structural Model	Combined Model	Combined Poll
Number of MGA Evolvables	NumEvo	**50**	**50**	**50**	**50**	**100**
Starting frequency of different motif component classes	**Gap**	**60%**	**60%**	**80%**	**40%**	**40%**
	IUPAC	**40%**	**0%**	**0%**	**20%**	**20%**
	WM	**0%**	**40%**	**0%**	**20%**	**20%**
	RangedCG	**0%**	**0%**	**20%**	**20%**	**20%**
Starting length of PosGenes	StartLen	75	75	75	75	75
Maximum DiscreteGene length	MaxDisc	**8**	**8**	**N/A**	**8**	**8**
Maximum Gap length	MaxGap	10	10	10	10	10
Maximum RangedContinuousGene length	**MaxRCG**	**N/A**	**N/A**	**12**	**12**	**12**
Fraction of parental Chromosomes that recombine	RecRate	0.2	0.2	0.2	0.2	0.2
Fraction of Chromosomes replaced by recombinants	RecRepl	0.6	0.6	0.6	0.6	0.6
BoundGene mutation rate	MutRateB	0.4	0.4	0.4	0.4	0.4
BoundGene maximum mutation amount	MutAmtB	100 nt	100 nt	100 nt	100 nt	100 nt
Fraction of MGA Chromosomes subject to BoundGene mutation	MutPropB	0.6	0.6	0.6	0.6	0.6
Rate of mutation affecting Gene type	MutRateT	0.2	0.2	0.2	0.2	0.2
Fraction of MGA Chromosomes subject to type mutation	MutPropT	0.6	0.6	0.6	0.6	0.6
Rate of type-conservative mutations	MutRateP	0.2	0.2	0.2	0.2	0.2
Maximum magnitude of type-conservative mutation	MutAmtP	0.2	0.2	0.2	0.2	0.2

Table 6-1, continued

Fraction of MGA Chromosomes subject to type-conservative mutation	MutPropP	0.6	0.6	0.6	0.6	0.6
Probability of changing the first Gene tested by the MGA Chromosome	MoveRate	0.2	0.2	0.2	0.2	0.2
Fraction of MGA Chromosomes subject to change in first Gene	MoveProp	0.4	0.4	0.4	0.4	0.4
Probability of joining or splitting adjacent motif components	JoinRate	0.6	0.6	0.6	0.6	0.6
Fraction of MGA Chromosomes subject to joining or splitting	JoinProp	0.2	0.2	0.2	0.2	0.2
Probability of swapping adjacent MGA Chromosome Genes	SwapRate	0.3	0.3	0.3	0.3	0.3
Fraction of MGA Chromosomes subject to Gene swapping	SwapProp	0.6	0.6	0.6	0.6	0.6
Rate of MGA Chromosome exchange between Evolvables	MigRate	**0.1**	**0.1**	**0.1**	**0.1**	**0.0**
Probability of converting a MotifGene to a gap	CrushRate	0.1	0.1	0.1	0.1	0.1
Fraction of MGA Chromosomes subject to gap replacement	CrushProp	0.6	0.6	0.6	0.6	0.6
Value of constant in the fitness function denominator	*c*	**N/A**	**N/A**	**N/A**	**N/A**	**0.01**
Contribution of specificity to fitness	K	**N/A**	**N/A**	**N/A**	**N/A**	**0.01**

Table 6-2: Prediction accuracy obtained from five MGA models of the CAP binding site. For each of the five types of model constructed, the percentage of positive cases correctly identified is shown, as well as the percentage of negative cases where cAMP/CAP binding sites were incorrectly identified. The overall accuracy is the difference between these two values. The training of the first four models differed only in the type of motif components that were allowed, while the fifth (Combined Poll) had no migration, a fitness function that assigned a more severe penalty to false positive predictions, and more MGA Chromosomes in each Evolvable (see Table 6-1 for details).

	IUPAC Model	Weight Matrix Model	Structural Model	Combined Model	Combined Poll Model
True Positive	100%	94.4%	94.4%	94.4%	72.2%
False Positive	13.9%	11.1%	3.3%	2.8%	2.2%
Accuracy	86.1%	83.3%	91.1%	91.8%	70.0%

Figure 6-1: The CAP consensus and best scoring models generated from five training runs performed to recognize and model the CAP dimer binding site. The notation is described in Chapter 5. Figure 6-1*a* shows the consensus binding site for CAP, with a number assigned to the base at each position. Figures 6-1*b* through 6-1*f* show the models obtained from the five training runs described in Table 6-1. For sequence-based motif components, the string of IUPAC and/or weight matrix positions are shown in order, with the minimum score required for a match shown below the motif component. IUPAC letters and their corresponding bases are shown in Table 1-1; a base that matches the appropriate IUPAC letter is assigned a score of 1.0, while a mismatch is assigned a score of 0.0. A weight matrix position shows the score associated with adenine, cytosine, guanine, and thymine in four rows. Gaps are shown with their associated length (L:) in nucleotides. Structural motif components also have an associated length, as well as an acceptable range (R:) for the averaged parameter value (See appendix A for structural mapping rules). The bottom row of a structural component indicates what percentage of all possible *n*-mers of length L are within the acceptable range R. In Figures 6-1*b* through 5-1*f*, the sites that are consistent with the CAP consensus are underlined, and the positions they match (if any) are indicated beneath the appropriate motif component. Where a motif component may match more than one position within the consensus, alternatives are indicated in parentheses.

a) Consensus:

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Base	**T**	**G**	**T**	**G**	**A**	**N**	**N**	**N**	**N**	**N**	**N**	**T**	**C**	**A**	**C**	**A**

b) IUPAC Model:

T	G	Y	D	A	**GAP**	H	**GAP**	H	D	W	H
					L: 9		**L: 0-1**				
5.0						**1.0**		**4.0**			
1	**2**	**3**	**4**	**5**	**6-14**	**15**		**16**			

c) Weight Matrix Model:

0.12	0.36	0.98	0.76	0.46	0.30	1	**GAP**	0.93	**GAP**	0.30
0	0.86	0.13	0.91	0.55	0.71	0.95		0		0.79
0	0.82	0.16	0	0	0.10	0.27	**L: 4-5**	0.88	**L: 4-5**	0.027
0.79	0.54	0.063	0	0.33	0.50	0.83		0		0.90
5.3								**0.86**		**0.76**
3	**4**	**5**	**6**	**7**	**8**	**9**	**10-13**	**14**		

Figure 6-1, continued

d) Structural Model:

BEND	GAP	BEND	LISSERFLEX	GAP	ROLL
L: 6-7	L: 4	L: 7	L: 4-5	L: 1	L: 2
R: 3.01/3.13		R: 2.85/2.93	R: 10.99/23.26		R: -4.74/0.59
6.34%		9.98%	35.77%		62.5%
1-6	7-10	11-16			

e) Combined Model:

H	MAJORDEP	ENTROPY	GAP	BEND	BRUNAKPROP	MAJORDIST	BEND
	L: 5	L: 3	L: 3	L: 4	L: 2-4	L: 3	L: 6
	R: 8.82/9.6	R: -23.77/-17.69		R: 2.58/2.93	R: -12.01/-11.93	R: 3.02/3.42	R: 2.88/3.91
1.0	70.51%	62.50%		35.94%	0.71%	75.00%	50.63%
					2-5 (3-6)	6-8 (7-9)	9-14 (10-15)

f) Combined Poll Model:

CLASH	GAP	SLIDE	FLEXOLSON	TWIST
L: 2	L: 8	L: 4	L: 3-4	L: 10
R: 0.43/1.78		R: 0.95/1.6	R: 2.25/2.5	R: 33.55/39.75
56.25%		1.95%	4.50%	99.17%
		1-4 (13-16)	5-7 (17-19)	

represented. The IUPAC motif shown in Figure 6-1*b* specifies each base in the first half of the CAP site (positions 1-5) with the TGYDA motif component, but most of the remaining site positions are in fact modeled with gaps, except for two H's at each end of the second half of the binding site. There is a slight favouring of weak bases at the end of this motif (DWH), which is consistent with extended consensus binding site of CAP. The weight matrix model shown in Figure 6-1*c* ignored the first two positions of the TGTGA motif, but the first three positions in the weight matrix clearly recognize the next three sites (consensus: TGA). All of the CAP sites in the positive set that deviate from the conserved TGA in positions 3 to 5 have either an A or a C at the sixth position, and this pattern is reflected in the high scores associated with those two bases at position 6. Any site that did not contain either TGA at positions 3-5 or A/C at position 6 would not be recognized by this model. Only a single residue from the second half of the motif (an A at position 14) was evident.

The final three models shown (Figure 6-1*d* to Figure 6-1*f*) rely entirely on structural representations. In each case, the structural features that imposed the greatest constraint (indicated by low percentage scores) recognized the conserved positions within the CAP binding site. In Figure 6-1*d*, both conserved halves of the motif were represented by the bending propensity of DNA. Though the motif is palindromic, each half was represented by a different range of values without overlap. According to the model shown here, the downstream half of the CAP binding site always has a smaller bend than the upstream half. Most of the hexamers and heptamers that contain TGTGA or TCACA, as well as the variants seen in the training set, have an average bend angle between 2.85 and 3.13.

In the 'combined' model of Figure 6-1*e*, the second half of the motif has a much less constrained bending range, but the first half is represented with propellor twist. The constraint on this component (0.71%) was the strongest seen in any of the three structural motifs shown in this figure, and only six DNA sequences (TGAC, TCAC, GTCA, GTGA, TGTC, and GACA) satisfy this requirement. While GTGA and GTCA are very common in positions 2-5 of the CAP motifs in the training set, TCAC and TGAC actually start at position 3 of their respective motifs. Thus, this feature occupies different (though overlapping) positions within different CAP sites in the training set. The bending

constraint that precedes the propellor twist component is not very restrictive, but it does exclude every tetramer that contains a TA dinucleotide step.

The motif in Figure 6-1*f* did not detect as many of the positive training set members as those in Figures 6-1*b* to 6-1*e*, due to the training conditions that prohibited migration and imposed heavy penalties for false positives. Only half of the CAP binding site is modeled with any degree of specificity. The range of the Slide motif allows only five of 256 tetramers: TGTG, CACA, CCCA, TGGG, and the palindromic TGCA. Of these, TGTG and CACA are identical to parts of the consensus, and TCGA is sometimes found in the first half of the training set motifs, while CCCA and TGGG are not found in any of the motifs. The flexibility constraint that follows the Slide motif component favours weak nucleotides either in the spacer or just downstream from the second half of the consensus.

In general, each of the five motif models appears to have detected at least half of the conserved pair of features. The structural model was the only one that placed similar weights on both halves of the motif, while the others appear to rely on a very specific model used to detect one half of each motif in the training set. This approach to modeling the CAP site is reasonable if only one half of the dimer would need to bind strongly to the DNA, thus stabilizing a weaker interaction between the other half of the dimer and its site.

Using Models to Detect Other Binding Sites

The models generated from all five training runs were searched against a nearly complete set of upstream regions from the chromosome of *E. coli* K12. Coding sequences that started within other coding sequences were not considered, reducing the number of available upstream regions from 4289 to 3888. These 3888 upstream regions included 12 of the 18 training set members (the others being plasmid-borne), and many upstream regions known to contain CAP binding sites (Stormo and Hartzell, 1989). Twenty-five genes known to be regulated by CAP were selected at random from RegulonDB (Salgado *et al.*, 2001) and used as a test set to determine the ability of the known motifs to detect novel patterns.

Table 6-3*a* shows the CAP site predictions obtained from the four models trained with migration (i.e., those shown in Figure 6-1*b* to 6-1*e*) on the training set members from the *E. coli* K12 genome. As shown in Table 6-2, patterns were found in all training set sequences by the IUPAC model, and the other three models missed one sequence each: the weight matrix model failed to detect the site associated with *tdcA*, while the structural and combined models both missed the *malE* site. The six plasmid-borne CAP-containing sequences from the training set are not shown here, but all six were matched by all four models.

The number of sites detected in each upstream region of the test set is shown for all four models in Table 6-3*b*. A CAP binding site was predicted in each test set upstream region by at least one of the four models. CAP sites were predicted in the *glpF* upstream region only by the structural model, while the *guaB* upstream region was matched by only the IUPAC model. Five upstream regions (of the *araF*, *epd*, *gcd*, *mtlA*, and *yiaJ* genes) contained sites that were identified by all four models. The structural motif matched the largest number of upstream sites (22/25 = 88%) in the test set, but also had the most hits throughout the genome, predicting a total of 5949 CAP sites in 3888 upstream regions for an average of 1.53 matches per upstream sequence. In contrast, the combined model found matches in 15 out of 25 test set upstream regions (60%), but predicted only 1516 CAP sites in the 3888 extracted upstream regions. This nearly fourfold difference in the number of genomic predictions is surprising, given the similar detection of false positives (3.3% for the structural model, 2.8% for the combined model) in the training set reported in Table 6-2. The weight matrix model found matches in fewer than half (12 / 25) of the test set upstream regions, with a prediction rate (3227 predicted sites = 0.83 matches per upstream region in the genome) more than double that of the combined model. Finally, the IUPAC model matched only one fewer test set upstream region than the structural model (21/25 = 84%), but its overall prediction rate was much lower (2566 predicted sites = 0.66 matches per upstream region in the genome).

Table 6-4 lists the 107 upstream regions that contained the most matches, averaged over all four models and ranked by the number of detected motifs. Only genes with annotated functions were included in this list; a further 84 putative coding sequences are not shown. Each member of the training set derived from the genome (listed in

Table 6-3: Summary of CAP binding sites detected by the best models from the IUPAC, weight matrix, structure, and combined training runs. The first column identifies the gene (names are consistent with the genome annotations available at www.ncbi.nlm.nih.gov and reported in Blattner *et al.* (1997)). In Table 6-3*a*, the presence or absence of a predicted motif in each training set sequence is indicated for each of the four models with a 'Yes' or 'No'. In Table 6-3*b*, the number (0 or more) of predicted cAMP/CAP binding sites is indicated for each upstream sequence in the test set (known to possess at least one motif) and for each of the four MGA models. The percentage of test set upstream regions in which at least one CAP binding site was predicted is shown, followed by the total number of predictions for ALL 3888 extracted upstream regions at the bottom of Table 6-3*b*.

a)

Training Set Member	IUPAC Model	Weight Matrix Model	Structural Model	Combined Model
araB	Yes	Yes	Yes	Yes
bglG	Yes	Yes	Yes	Yes
crp	Yes	Yes	Yes	Yes
cya	Yes	Yes	Yes	Yes
deoC	Yes	Yes	Yes	Yes
galE	Yes	Yes	Yes	Yes
ilvB	Yes	Yes	Yes	Yes
lacZ	Yes	Yes	Yes	Yes
malE	Yes	Yes	No	No
ompA	Yes	Yes	Yes	Yes
tnaA	Yes	Yes	Yes	Yes
tdcA	Yes	No	Yes	Yes

b)

Test Set Member	IUPAC Model	Weight Matrix Model	Structural Model	Combined Model
araF	1	1	3	2
araJ	2	0	1	0
aspA	1	0	1	0
caiF	2	2	1	0
cytR	2	0	3	1
dadA	2	0	2	0
epd	1	1	1	1
fixA	4	0	2	1
gcd	1	2	1	1
glgS	3	0	1	0
glpF	0	0	2	0
guaB	1	0	0	0
hpt	2	0	2	1
mglB	1	0	1	1
mlc	0	1	1	1
mtlA	2	4	2	1
nupC	1	0	3	1
osmY	3	0	2	2
ppiA	2	1	0	1
ptsI	1	2	0	1
rhaT	0	1	2	0
rpoH	0	1	1	0
sohB	2	1	2	0
uhpT	1	0	3	1
yiaJ	1	2	2	1
% matched	**84%**	**48%**	**88%**	**60%**
Total (All Upstream Regions)	**2556**	**3227**	**5949**	**1516**

Table 6-4: Upstream regions with the largest number of detected cAMP/CAP sites. All 107 upstream regions with an average of 1.75 or more identified motifs across the four different models (I = IUPAC, W = weight matrix, S = structural, C = combined), with a function assigned by the annotators (Blattner *et al.*, 1997), are listed below. The gene name associated with the upstream region is indicated in the first column, with the annotated function indicated in the second column. Upstream regions with representation in the training set are indicated in boldface. Missing training set members: *crp* = 1.5, *lacZ* = 1.5.

Name	Gene Function	Number of sites detected				Mean
		I	W	S	C	
uxuA	**D-mannonate hydrolase**	**4**	**3**	**4**	**2**	**3.25**
amn	AMP nucleosidase	1	2	5	2	2.5
galF	UTP-glucose-1-phosphate uridylyltransferase	2	0	7	1	2.5
galP	galactose-proton symport	3	2	4	1	2.5
galR	galactose operon repressor	4	3	2	1	2.5
gppA	guanosine-5'-triphosphate 3'-diphosphate	2	3	4	1	2.5
ilvBN	**operon leader peptide**	**4**	**2**	**3**	**1**	**2.5**
malE	**periplasmic maltose-binding protein**	**3**	**3**	**3**	**1**	**2.5**
xthA	exodeoxyribonuclease III	3	2	4	1	2.5
araB	**L-ribulokinase**	**2**	**2**	**3**	**2**	**2.25**
ascG	asc operon repressor protein	2	1	3	3	2.25
csiE	stationary phase inducible protein CsiE	2	1	5	1	2.25
cyaA	**adenylate cyclase**	**2**	**2**	**3**	**2**	**2.25**
deoC	**deoxyribose-phosphate aldolase**	**2**	**1**	**3**	**3**	**2.25**
dppA	periplasmic dipeptide transport protein	3	2	3	1	2.25
fecA	iron(III) dicitrate transport protein FecA	2	3	2	2	2.25
fucA	fuculose-1-phosphate aldolase	3	2	1	3	2.25
gntP	gluconate permease 3	0	3	4	2	2.25
malK	cytoplasmic membrane protein for maltose uptake	2	1	3	3	2.25
mtlA	mannitol-specific enzyme II of	2	4	2	1	2.25
narQ	nitrate/nitrite sensor protein narQ	1	2	4	2	2.25
nuoA	NADH dehydrogenase I chain A	3	0	3	3	2.25
ompA	**outer membrane protein A**	**2**	**2**	**3**	**2**	**2.25**
tnaA	**tryptophanase**	**3**	**1**	**3**	**2**	**2.25**
appY	M5 polypeptide	3	0	4	1	2
argT	lysine-arginine-ornithine-binding periplasmic	4	0	3	1	2
asnC	regulatory protein	2	1	3	2	2
aspC	aspartate aminotransferase	1	2	2	3	2
atpI	ATP synthase subunit	2	1	3	2	2
cbpA	curved DNA-binding protein	1	2	4	1	2
csgG	assembly /transport component	0	4	3	1	2
dapA	dihydrodipicolinate synthase	1	2	3	2	2
fsr	fosmidomycin resistance protein	3	1	4	0	2
glpT	glycerol-3-phosphatase transporter	2	2	4	0	2
grpE	heat shock protein grpE (heat shock protein)	1	2	3	2	2
kdpD	sensor protein KdpD	0	2	6	0	2
nuoF	NADH dehydrogenase I chain F	1	2	3	2	2
ompT	protease VII precursor	2	3	3	0	2
proP	proline/betaine transporter (proline porter II)	3	2	3	0	2
pyrB	aspartate carbomoyltransferase catalytic	1	3	2	2	2
rbsA	high affinity ribose transport protein	0	5	2	1	2
rfaK	lipopolysaccharide 1 2-n-	5	0	3	0	2
rpoE	RNA polymerase sigma-E factor (sigma-24)	2	1	4	1	2
tesB	acyl-coA thioesterase II	1	2	5	0	2
tktA	transketolase	2	0	5	1	2
tsx	nucleoside-specific channel-forming protein Tsx	5	1	1	1	2
uvrC	excinuclease ABC subunit C	2	3	2	1	2

Table 6-4, continued

araF	L-arabinose-binding periplasmic protein	1	1	3	2	1.75
arcA	aerobic respiration control protein ArcA	4	2	1	0	1.75
bglX	periplasmic beta-glucosidase precursor	1	1	3	2	1.75
bioA	adenosylmethionine-8-amino-7-oxononanoate	2	2	2	1	1.75
coaA	pantothenate kinase	3	2	2	0	1.75
cydA	cytochrome d ubiquinol oxidase subunit I	1	2	4	0	1.75
cyoA	cytochrome o ubiquinol oxidase subunit II	4	2	1	0	1.75
endA	endonuclease I	2	1	4	0	1.75
fdoG	formate dehydrogenase-O alpha subunit	1	3	3	0	1.75
ffh	signal recognition particle protein	1	1	3	2	1.75
fixA	FixA protein	4	0	2	1	1.75
flgB	putative flagellar basal-body rod protein FlgB	1	2	4	0	1.75
flhC	flagellar transcriptional activator	1	2	4	0	1.75
fnr	fumarate and nitrate reduction regulatory	3	2	1	1	1.75
fur	ferric uptake regulation protein	2	1	2	2	1.75
galE	**UDP-glucose 4-epimerase**	**2**	**2**	**2**	**1**	**1.75**
gcvT	aminomethyltransferase	0	1	4	2	1.75
glcC	glc operon transcriptional activator	2	3	1	1	1.75
glcD	glycolate oxidase subunit GlcD	2	2	3	0	1.75
glnA	glutamine synthetase	0	2	4	1	1.75
gmk	5'guanylate kinase	2	1	3	1	1.75
hdhA	7-alpha-hydroxysteroid dehydrogenase	1	2	4	0	1.75
himD	integration host factor beta-subunit (IHF-beta)	3	1	2	1	1.75
hrpB	ATP-dependent helicase HrpB	0	2	5	0	1.75
ilvB	acetohydroxy acid synthase I small subunit	3	1	2	1	1.75
kdul	5-keto-4-deoxyuronate isomerase	0	1	5	1	1.75
lig	DNA ligase	1	2	2	2	1.75
lpxC	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine	2	2	3	0	1.75
mcrC	McrC protein	0	2	5	0	1.75
mdoH	periplasmic glucans biosynthesis protein MdoH	0	1	5	1	1.75
metB	cystathionine gamma-synthase	0	2	3	2	1.75
murI	glutamate racemase	2	2	1	2	1.75
nac	nitrogen assimilation control protein	2	2	2	1	1.75
nagB	glucosamine-6-phosphate isomerase	4	2	0	1	1.75
narP	nitrate/nitrite response regulator protein NarP	2	1	4	0	1.75
ndk	nucleoside diphosphate kinase (ndk)	3	0	3	1	1.75
oppC	oligopeptide transport system permease protein	2	1	3	1	1.75
osmY	periplasmic protein	3	0	2	2	1.75
pdhR	pyruvate dehydrogenase complex repressor	2	2	2	1	1.75
pgk	phosphoglycerate kinase	3	0	4	0	1.75
pgm	phosphoglucomutase	0	1	6	0	1.75
phpB	phosphohistidine protein	0	4	2	1	1.75
pldA	detergent-resistant phospholipase A	2	2	3	0	1.75
potF	putrescine-binding periplasmic protein	4	1	2	0	1.75
ptsN	enzyme IIANtr	1	3	3	0	1.75
recD	exonuclease V alpha-subunit	2	2	3	0	1.75
recQ	DNA-dependent ATPase DNA helicase	1	4	2	0	1.75
rhsD	RhsD protein precursor	5	0	2	0	1.75
rnhA	ribonuclease H	1	3	1	2	1.75
rob	right origin-binding protein	3	1	2	1	1.75
rpsF	30S ribosomal subunit protein S6	1	1	4	1	1.75

Table 6-4, continued

rpsQ	30S ribosomal subunit protein S17	1	1	3	2	1.75
ruvA	Holliday junction DNA helicase RuvA	0	2	4	1	1.75
sbmA	SbmA protein	2	1	3	1	1.75
sodA	manganese superoxide dismutase	1	2	3	1	1.75
tdcA	**tdcABC operon transcriptional activator**	**3**	**1**	**2**	**1**	**1.75**
torA	trimethylamine-N-oxide reductase precursor	1	2	4	0	1.75
tpx	thiol peroxidase	2	2	3	0	1.75
truB	tRNA pseudouridine 55 synthase (psi55 synthase)	1	2	3	1	1.75
zwf	glucose 6-phosphate 1 dehydrogenase	1	1	3	2	1.75

6-3*a*) was matched at least 1.5 times on average, but *lacZ* and *crp* are not included in this table because the >100 genes with an average of 1.5 matches are not shown. Only four of 25 test set members (*araF, fixA, mtlA,* and *osmY*) are in this list, and two more (*cytR* and *yiaJ*) have an average of 1.5 matches per model. However, the list in Table 6-4 contains a number of plausible targets for CAP control: in addition to the *araF* gene known to contain a CAP binding site, the trained models predicted several CAP sites upstream of genes involved in the transport and metabolism of sugars such as galactose (*galF, galP, galR*) and fuculose (*fucA*). While few CAP sites were predicted upstream of the gene encoding σ^{32} (*rpoH*), the heat shock proteins *grpE* and *rpoE* each had an average of 2 sites predicted per model. Genes from other systems known to be regulated by CAP were also identified, including iron uptake (*fecA, fur*), flagellar synthesis (*flgB, flhC*) and drug resistance (*fsr*). Several other genes involved in sugar metabolism also possess a high average number of predicted CAP sites (*gntP, glpT, rbsA, bglX, nagB, pgm, zwf*). Several genes, including *rpsF* and *rpsQ* which encode ribosomal small subunit proteins, have not been associated with CAP to date, and may thus be less likely targets for the regulatory protein. The presence of such genes underscores the need for experimental validation of potential targets of CAP or other regulatory proteins.

Tables 6-5*a* and 6-5*b* summarize the motifs found by the best 25 of 100 MGA Evolvables trained without migration in the 'Combined Poll' experiment. The number of predicted motifs are counted two ways: the gross number of predictions sums the entire number of predictions made by each of the 25 models, while the model count includes each model only once, even if a given MGA Chromosome makes multiple motif predictions within a single upstream region. Within the training set, the number of motifs matching each upstream sequence ranged from a maximum of 23/25 for *malE* down to 5/25 for *bglG*. While most (10/12) training set upstream regions were within the top 250 out of all 3888 when ranked on model count, *galE* (matched by 7 models, tied for 1327th to 1918th position) and *bglG* (matched by 5 models, tied for 2554th to 3127th position) were not well represented. These rankings improved dramatically when gross count was considered in place of model count, with *galE* and *bglG* rising to 68th to 106th and 385th to 573rd positions, respectively. The few models that match these two genes must detect several binding sites each to increase their ranking. In fact, all members of the training set

Table 6-5: Summary of cAMP/CRP binding sites detected by the 25 best models generated in the 'combined poll' training run. Table 6-5*a* shows the sites detected in the training set sequences, while Table 6-5*b* shows the sites detected in a set of sequences from RegulonDB known to be regulated by CAP. The gene names associated with upstream regions in the training and test set are as in Table 6-3. The second column identifies how many of the 25 models identified at least one cAMP/CRP binding site in the 400 nt upstream region of the gene. The fourth column identifies the total number of predicted sites that were obtained from all 25 models. The third and fifth columns show the rank of each training and test set member, respectively, when compared to all 3888 upstream regions extracted from the *E. coli* K12 genome. The third column counts each model either zero or one times, ignoring multiple hits to a given upstream region from the same MGA Chromosome. In the fifth column, upstream regions are ranked on the total number of matches with all models, thus allowing multiple matches from a single MGA Chromosome. The rankings are shown with gray bars whose horizontal positions correspond to the ranking among all 3888 upstream regions, with the highest rank (largest number of model hits) at the left end, and the lowest rank at the right, and tied rankings indicated by the width of the bar. The number of upstream regions considered (3888) is less than the total annotated number of genes (4289) because coding regions that began within other coding regions were not considered.

a)

Training Set Member	Model Count		Gross Count	
	# of models recognized / 25	Rank / 3888	# of predicted CAP binding sites	Rank / 3888
araB	13		19	
bglG	5		12	
crp	11		22	
cya	12		18	
deoC	15		34	
galE	7		16	
ilvB	13		23	
lacZ	13		20	
malE	23		36	
ompA	11		15	
tnaA	15		22	
tdcA	10		16	

b)

Test Set Member	Model Count		Gross Count	
	# of models recognized / 25	Rank / 3888	# of predicted CAP binding sites	Rank / 3888
araF	6		6	
araJ	8		8	
aspA	7		8	
caiF	5		8	
cytR	8		16	
dadA	7		7	
epd	8		7	
fixA	13		11	
gcd	3		7	
glgS	9		9	
glpF	11		15	
guaB	5		10	
hpt	7		11	
mglB	9		8	
mlc	10		13	
mtlA	17		21	
nupC	7		7	
osmY	6		14	
ppiA	8		13	
ptsI	9		11	
rhaT	6		9	
rpoH	5		8	
sohB	4		9	
uhpT	5		5	
yiaJ	6		10	

(with the exception of first-ranked *malE*) increased in rank from the model count to the gross count.

For the most part, the performance of the trained models on the test set (Figure 6-5*b*) failed to distinguish the test set members from the bulk of upstream sequences in the genome. While a difference of two or three models matching an upstream sequence (*e.g.* 6/25 versus 8/25) may yield a dramatic change in rank without being very significant, almost half of upstream sequences in the test set were matched by fewer than ¼ of all trained models. Only 8 upstream regions ever achieved a rank above 500, and only 2 of these achieved better than this rank for both the model count and the gross count (*mtlA*: model rank 4/3888, gross rank 7/3888; *glpF*: model rank 128/3888, gross rank 115/3888). This failure to emphasize the test set sequences suggests either that the fitness function designed to limit the number of false positives impeded model learning, or that the training set sequences were not sufficiently representative of the range of CAP sites to train models capable of generalization. Since the combined and IUPAC models in particular were able to detect relevant patterns in most of the test set upstream regions while maintaining a reasonable degree of stringency, the limitation seen here is more likely due to the relatively low accuracy of the 25 models shown here. Note that since this method did not perform well in detecting members of the test set, the upstream regions containing the largest number of predicted CAP sites were not analyzed, as was performed for the four migration-enabled models in Table 6-4.

Hypothesis Testing:
Analysis of IHF Binding Sites

Overview

Integration host factor (IHF) is a nucleoid protein in *E. coli* that binds DNA and induces drastic bends of up to 180° in the helical structure (Rice *et al.*, 1996). IHF is important in functions such as DNA replication, site-specific recombination, and DNA packaging (Dhavan *et al.*, 2002). Within the context of transcription, IHF can bring

upstream activator binding sites (sometimes called enhancers) into close proximity with the promoter to allow activator protein interactions with RNA polymerase (Dworkin *et al.*, 1998). IHF is a heterodimeric protein, with each subunit containing a β-sheet that interacts with and unwinds the helix at a single base step, in each case yielding a local bend of ~90° (Travers, 1997).

While a consensus sequence, AAAAAA(N_8)TATCAA(N_4)TTGC, has been determined for IHF (Bewley *et al.*, 1998), the key property of DNA that allows IHF binding and action is the deformation energy of the two sites where bending is induced (Steffen *et al.*, 2002). This dependence on conformational mobility explains the conservation of deformable steps, particularly CA/TG at the unwound sites, since pyrimidine/purine steps tend to be the most deformable (el Hassan and Calladine, 1996). Further interactions occur between the protein and a conserved A-tract just upstream of the 180° bend in the DNA (Bewley *et al.*, 1998), and with a conserved TTG just downstream (Ellenberger and Landy, 1997).

In this experiment, IHF sites were obtained via a query to RegulonDB (Salgado *et al.*, 2001). Of the 64 sites stored within the database, 42 had been verified through binding assays. The other 22 were identified either through computational prediction or an observed effect of the protein on transcription *in vivo*, without direct evidence for binding. Some of the 42 verified sites only showed the central ~10 nt corresponding to the TATCAAN part of the site, so the entire set of binding sites were extracted from the *E. coli* K12 genome with enough flanking nucleotides to yield 30 nt of sequence for each site. These 42 sequences comprised the positive set, and the negative set was generated as with CAP above, with five shuffled copies of each positive set sequence.

Model Construction

Three separate IHF models were constructed, with a different range of permissible motif types in each. One model was permitted the use of IUPAC and weight matrix descriptions, yielding motif models constructed only from sequence and gaps. The second model was allowed to use the IUPAC and weight matrix descriptions as well, but was also permitted the use of a single structural category. Since deformation energy is

important in IHF interactions with DNA (Steffen *et al.*, 2002), the dinucleotide flexibility values from Olson *et al.* (1998) used to determine the deformation energy of a binding site were included in the training of the second model. The third and final model included IUPAC and weight matrix descriptions, and could use the entire set of defined structural values in Appendix A. The parameters for each MGA training run are shown in Table 6-6. Migration was enabled in all three training runs, and the fitness function was the same as in the first four CAP detection experiments from the **Constrained Models** section above.

The classification accuracy of the best MGA Chromosome obtained from each of the three experiments is indicated in Table 6-7. All three models achieved prediction accuracies better than 60%, and the rate of false positive classification (13.3 – 14.3%) was consistent among all three models. However, the model that was constrained to use only the FLEXOLSON structural relationship identified the greatest number of positive set members (39 out of 42), for a total classification accuracy of 79.6%. The combined model, which could draw from all available structural categories, had an overall classification accuracy of 71.4%.

The three highest scoring motif models are shown in Figure 6-2. The IUPAC/weight matrix model (Figure 6-2*b*) used only IUPAC degenerate designations to model conserved patterns within the IHF binding sites. The motif component that requires a perfect match to VAHCHNBN appears to recognize variants of the central TATCAA consensus in IHF, but the initial V would prevent the consensus itself from matching. There appears to be little or no recognition of the upstream A-tract and the downstream TTG. The model in Figure 6-2*c* appears to recognize the central TATCAA motif as well. At the upstream end, two consecutive non-A bases are required, followed by a dinucleotide step with flexibility between 2.65 and 2.9, which can only be satisfied by an AA or TT step. This condition is often met with either TTAA or TTTT in the training set. After a gap of ~0.5 helical turns is another strong flexibility constraint of length 5. The flexibility of the pentanucleotide ATCAA and many single-base variants is within the required range. The downstream constraints are much weaker, and exclude only the least (such as AT and AG/CT) and most (CG) flexible step types. While the downstream TTG would be recognized by this model, many other less flexible sequences

Table 6-6: Summary of parameter values used in the three MGA training runs performed on IHF binding sites. Detailed descriptions of the relevant MGA functions can be found in the **Design and Implementation** section of Chapter 5. Abbreviated parameter names are shown in the second column, and parameters whose values were different in one or all of the experiments are indicated with boldface text.

Parameter Description	Parameter Name	Sequence (IUPAC + WM) Model	Sequence + FLEXOLSON Model	Sequence + All Structure Model
Number of MGA Evolvables	NumEvo	50	50	50
Starting frequency of different motif component classes	**Gap**	**60%**	**40%**	**40%**
	IUPAC	20%	20%	20%
	WM	20%	20%	20%
	RangedCG	**0%**	**20%**	**20%**
Starting length of PosGenes	StartLen	10	10	10
Maximum DiscreteGene length	MaxDisc	10	10	10
Maximum Gap length	MaxGap	10	10	10
Maximum RangedContinuousGene length	**MaxRCG**	**N/A**	**10**	**10**
Fraction of parental Chromosomes that recombine	RecRate	0.2	0.2	0.2
Fraction of Chromosomes replaced by recombinants	RecRepl	0.6	0.6	0.6
BoundGene mutation rate	MutRateB	0.4	0.4	0.4
BoundGene maximum mutation amount	MutAmtB	100 nt	100 nt	100 nt
Fraction of MGA Chromosomes subject to BoundGene mutation	MutPropB	0.6	0.6	0.6
Rate of mutation affecting Gene type	MutRateT	0.2	0.2	0.2
Fraction of MGA Chromosomes subject to type mutation	MutPropT	0.6	0.6	0.6

Table 6-6, continued

Rate of type-conservative mutations	MutRateP	0.2	0.2	0.2
Maximum magnitude of type-conservative mutation	MutAmtP	0.2	0.2	0.2
Fraction of MGA Chromosomes subject to type-conservative mutation	MutPropP	0.6	0.6	0.6
Probability of changing the first Gene tested by the MGA Chromosome	MoveRate	0.2	0.2	0.2
Fraction of MGA Chromosomes subject to change in first Gene	MoveProp	0.4	0.4	0.4
Probability of joining or splitting adjacent motif components	JoinRate	0.6	0.6	0.6
Fraction of MGA Chromosomes subject to joining or splitting	JoinProp	0.2	0.2	0.2
Probability of swapping adjacent MGA Chromosome Genes	SwapRate	0.3	0.3	0.3
Fraction of MGA Chromosomes subject to Gene swapping	SwapProp	0.6	0.6	0.6
Rate of MGA Chromosome exchange between Evolvables	MigRate	0.1	0.1	0.1
Probability of converting a MotifGene to a gap	CrushRate	0.1	0.1	0.1
Fraction of MGA Chromosomes subject to gap replacement	CrushProp	0.6	0.6	0.6
Value of constant in the fitness function denominator	*c*	N/A	N/A	N/A
Contribution of specificity to fitness	K	N/A	N/A	N/A

Table 6-7: Classification accuracy obtained from three MGA models of the IHF binding site. For each of the three types of model constructed, the percentage of positive cases correctly identified is shown, as well as the percentage of negative cases where IHF binding sites were incorrectly identified. The overall accuracy is the difference between these two values.

	Sequence (IUPAC + WM) Model	Sequence + FLEXOLSON Model	Sequence + All Structure Model
True Positive	76.2%	92.9%	85.7%
False Positive	13.8%	13.3%	14.3%
Accuracy	62.4%	79.6%	71.4%

Figure 6-2: The IHF binding site consensus (5-2*a*) and the three best scoring models (*b*, sequence only; *c*, sequence and the FLEXOLSON structural relationship; *d*, sequence and all structural relationships) trained to recognize the IHF binding site. For sequence-based motif components, the string of IUPAC and/or weight matrix positions are shown in order, with the minimum score required for a match shown below the motif component. IUPAC letters and their corresponding bases are shown in Table 1-1; a base that matches the appropriate IUPAC letter is assigned a score of 1.0, while a mismatch is assigned a score of 0.0. A weight matrix position shows the score associated with adenine, cytosine, guanine, and thymine in four rows. Gaps are shown with their associated length (L:) in nucleotides. Structural motif components also have an associated length, as well as an acceptable range (R:) for the averaged parameter value (See appendix A for structural mapping rules). The bottom row of a structural component indicates what percentage of all possible *n*-mers of length L are within the acceptable range R. In Figures 6-2*b* through 6-2*d*, the sites that are consistent with the IHF consensus are underlined, and the positions they match (if any) are indicated beneath the appropriate motif component.

a) Consensus:

Position	1	2	3	4	5	6	7-14	15	16	17	18	19	20	21-24	25	26	27	28
Base	**A**	**A**	**A**	**A**	**A**	**A**	**N**	**T**	**A**	**T**	**C**	**A**	**A**	**N**	**T**	**T**	**G**	**C**

b) Sequence:

D	**GAP**	V	A	H	C	H	N	B	N	D	H	M	H
	L: 0-2												
1.0		**8.0**								**3.0**			
		15-22								**23-25**			

c) Sequence + FLEXOLSON:

B	B	**FLEXOLSON**	**GAP**	**FLEXOLSON**	**FLEXOLSON**	**FLEXOLSON**
		L:2	**L: 4-7**	**L: 5**	**L: 3**	**L: 2**
		R: 2.65/2.9		**R: 4.34-4.71**	**R: 2.21/7.95**	**R: 2.19/9.80**
2.0		**12.5%**		**12.9%**	**87.5%**	**75.0%**
				16-20	**21-23**	**24-25**

Figure 6-2, continued

d) Sequence + All Structure:

BRUNAKSTACK	GAP	FLEXOLSON	CLASH	GAP	NUCLEOSOME
L: 2	L: 5	L: 4	L: 2	L: 0-1	L: 5
R: -6.32/-4.86		R: 4.77/5.53	R: 0/2.52		R: 9.9/14.06
12.5%		16.8%	93.8%		60.7%
		17-20	21-22		23-27

would be accepted as well.

The model in Figure 6-2*d* is an interesting contrast to the model based on the FLEXOLSON relationship. The first component restricts the dinucleotide stacking energy (Baldi *et al.*, 1998) to a range between –6.32 and –4.86, which, like the first structural constraint in the previous model, excludes all dinucleotides but AA and TT. After a gap of 5 nt, a flexibility constraint is seen, but the range of this feature does not overlap with that of the corresponding motif in the FLEXOLSON model. This shorter feature recognizes ATCA from the consensus with the higher mean flexibility value due to the exclusion of the inflexible AA step. As with the previous model, there is only a weak constraint imposed by the remaining model features. In both of the motifs that include DNA structure, there is a requirement for rigid AA or TT steps upstream of the central TATCAA, but the spacing within the models is less than the spacing from the upstream A-tract to the central conserved position.

New Motifs:

Pattern Detection in Upstream Regions of Co-expressed Genes

Overview – Alternate sigma factors

Bacteria typically have a single vegetative sigma factor that associates with the RNA polymerase core enzyme to allow recognition of most promoters. However, stressful physiological conditions and/or different growth stages can require the expression of different sets of genes. In some cases, this change in expression is effected through regulons under the control of alternative sigma factors (Wösten, 1998). This set of experiments will search for patterns upstream of genes regulated by two of these alternative sigma factors: σ^{32} from *Escherichia coli* K12, and σ^{B} from *Bacillus subtilis* 168.

The σ^{32} protein of *E. coli* is the major sigma factor required for the heat-shock response. The control of *rpoH*, the gene that encodes σ^{32}, is quite complex, with four separate promoters: three under the control of vegetative σ^{70}, and one responsive to the extreme heat-shock σ^{E} protein (Gross, 1996). While translational control is important in

the regulation of σ^{32} expression, mRNA production is under several forms of control, including heat (via σ^{E} and a heat-sensitive σ^{70} promoter), replication (via the DnaA protein), and nucleoside metabolism (via CytR and CAP) (Kallipolitis *et al.*, 1998). The expression of over 20 proteins is under the control of σ^{32}, including the proteases Lon, ClpP, and ClpX, and chaperones such as DnaK, GroEL, and GroES (Rosen and Ron, 2002).

The σ^{32} promoter is similar in organization to the more familiar σ^{70}, with –10 and –35 consensi, and a tendency for runs of As or Ts in the –45 to –50 region. However, the nucleotide composition of the σ^{32} –35 and –10 boxes is different from that of σ^{70}: the –35 consensus is CTTGAAA, and the –10 is CCCATNT, in contrast to TTGACA and TATAAT for *E. coli* σ^{70} (Gross, 1996). Also unlike the σ^{70} consensus, some σ^{32} promoters such as promoter 1 of *htpG* actually match the entire consensus perfectly.

To generate the positive training set, the upstream regions of 16 genes known to have σ^{32} promoters (Gross, 1996) were extracted. Four hundred nucleotides of upstream sequence were extracted for each gene. Each sequence was again shuffled 5 times to yield a negative set of size 80.

The Gram-positive bacterium *Bacillus subtilis* has a large set of stress induced pathways that are ultimately under the control of the alternative sigma factor σ^{B}. This protein recognizes and permits transcriptional activation of genes in response to stresses such as ethanol, high salt, heat, and acid, as well as nutrient starvation (Hecker and Volker, 1998; Hecker *et al.*, 1996). The transduction of physiological signals leading to this response is complex and not completely understood, but control of σ^{B} appears to be effected mainly through an anti-sigma factor protein, RsbW, which binds σ^{B} when the cell is not stressed. Stress signals lead to the activation of the RsbV protein, which preferentially binds RsbW and frees σ^{B} to activate transcription (Benson and Haldenwang, 1993).

The consensi of several different promoter types were searched against the whole genome of *Bacillus subtilis* 168 (Huang and Hellman, 1998; Huang *et al.*, 1999). Petersohn *et al.* (1999) searched the genome with the σ^{B} consensus

GTTTWW(N_{12-15})GGGWAW and identified a number of known promoters as well as new ones, which were then investigated for σ^B dependence experimentally. For the present analysis, the positive set consisted of 400 nt upstream regions extracted from genes known to be under the control of σ^B promoters. Twenty-two genes were selected from a list presented in Helmann *et al.* (2001). As with the previous experiments, the negative set contained 5 shuffled copies of each upstream region from the positive set, for a total of 110 negative set sequences.

Model Construction

Two models were constructed for each set of sequences: a sequence model that was constrained to use only IUPAC and weight matrix representations, and a combined model that could use any combination of sequence and structural indices. Table 6-8 lists the MGA parameter values for these two types of experiment. The maximum gap size was set to 20 nt to allow modeling of the spacer region between conserved promoter elements. These runs used the same fitness function as the IHF and CAP runs to rank the MGA Chromosomes.

Table 6-9 shows the classification accuracy of the two models trained on the *E. coli* K12 promoter set. Both models correctly classified the same number (15/16 = 93.8%) of positive set sequences, but the model based on sequence and structure misclassified only 2.5% of the negative set sequences, versus a 5.0% misclassification rate for the model based on sequence alone. The motif models that yielded these levels of classification accuracy are shown in Figure 6-3. The sequence model in Figure 6-3*b* resembles the consensus in that is rich in cytosines, but the consensus does not align well with the modeled motif. The guanine residues in the modeled motif do not correspond to the correct consensus positions. The structural model in Figure 6-3*c* presents a similar conundrum: while the size of the gap (13 nt) coincides with the spacing between many of the paired σ^{32} consensus elements, the downstream component of the motif model demands a guanine, which does not exist in the consensus. However, the melting temperature constraints match several –10 sequences, including the consensus CCCCAT which is assigned an average melting temperature of 73.9° C by the structural rules.

Table 6-8: Summary of parameter values used in training MGA Chromosomes to detect patterns in 400 nucleotide upstream sequences containing either *Escherichia coli* K12 σ^{32} or *Bacillus subtilis* 168 σ^{B} promoters. Two types of run (sequence, sequence + structure) were performed for each set of upstream sequences. Detailed descriptions of the relevant MGA functions can be found in the **Design and Implementation** section of Chapter 5. Abbreviated names of important parameters are shown in the second column, and parameters whose values were different in one or all of the experiments are indicated with boldface text.

Parameter Description	Parameter Name	Sequence Model	Sequence + Structure Model
Number of MGA Evolvables	NumEvo	50	50
Starting frequency of different motif component classes	**Gap**	**60%**	**40%**
	IUPAC	20%	20%
	WM	20%	20%
	RangedCG	**0%**	**20%**
Starting length of PosGenes	StartLen	300	300
Maximum DiscreteGene length	MaxDisc	12	12
Maximum Gap length	MaxGap	20	20
Maximum RangedContinuousGene length	**MaxRCG**	**N/A**	**12**
Fraction of parental Chromosomes that recombine	RecRate	0.2	0.2
Fraction of Chromosomes replaced by recombinants	RecRepl	0.6	0.6
BoundGene mutation rate	MutRateB	0.4	0.4
BoundGene maximum mutation amount	MutAmtB	100 nt	100 nt
Fraction of MGA Chromosomes subject to BoundGene mutation	MutPropB	0.6	0.6
Rate of mutation affecting Gene type	MutRateT	0.2	0.2

Table 6-8, continued

Fraction of MGA Chromosomes subject to type mutation	MutPropT	0.6	0.6
Rate of type-conservative mutations	MutRateP	0.2	0.2
Maximum magnitude of type-conservative mutation	MutAmtP	0.2	0.2
Fraction of MGA Chromosomes subject to type-conservative mutation	MutPropP	0.6	0.6
Probability of changing the first Gene tested by the MGA Chromosome	MoveRate	0.2	0.2
Fraction of MGA Chromosomes subject to change in first Gene	MoveProp	0.4	0.4
Probability of joining or splitting adjacent motif components	JoinRate	0.6	0.6
Fraction of MGA Chromosomes subject to joining or splitting	JoinProp	0.2	0.2
Probability of swapping adjacent MGA Chromosome Genes	SwapRate	0.3	0.3
Fraction of MGA Chromosomes subject to Gene swapping	SwapProp	0.6	0.6
Rate of MGA Chromosome exchange between Evolvables	MigRate	0.1	0.1
Probability of converting a MotifGene to a gap	CrushRate	0.1	0.1
Fraction of MGA Chromosomes subject to gap replacement	CrushProp	0.6	0.6
Value of constant in the fitness function denominator	*c*	N/A	N/A
Contribution of specificity to fitness	K	N/A	N/A

Table 6-9: Classification accuracy obtained from two MGA models trained on 400 nt upstream regions containing *E. coli* K12 σ^{32} promoters. For both models, the percentage of positive cases correctly identified is shown, as well as the percentage of negative cases where patterns were incorrectly identified. The overall accuracy is the difference between these two values.

	Sequence Model	Sequence + Structure Model
True Positive	93.8%	93.8%
False Positive	5.0%	2.5%
Accuracy	88.8%	91.3%

Figure 6-3: The best scoring models trained on upstream sequences containing *E. coli* K12 σ^{32} promoters. The promoter consensus is shown in Figure 6-3*a*, with the variable length spacer region represented by an asterisk. Figure 6-3*b* shows a model trained with sequence components only, while Figure 6-3*c* shows the best model trained with sequence and structural components. The notation is described in Chapter 5. For sequence-based motif components, the string of IUPAC and/or weight matrix positions are shown in order, with the minimum score required for a match shown below the motif component. IUPAC letters and their corresponding bases are shown in Table 1-1; a base that matches the appropriate IUPAC letter is assigned a score of 1.0, while a mismatch is assigned a score of 0.0. A weight matrix position shows the score associated with adenine, cytosine, guanine, and thymine in four rows. Gaps are shown with their associated length (L:) in nucleotides. Structural motif components also have an associated length, as well as an acceptable range (R:) for the averaged parameter value (See appendix A for structural mapping rules). The bottom row of a structural component indicates what percentage of all possible *n*-mers of length L are within the acceptable range R. In Figures 6-3*b* and 6-3*c*, the sites that are consistent with the σ^{32} consensus are underlined, and the positions they match (if any) are indicated beneath the appropriate motif component.

a) Consensus:

Position	A1	A2	A3	A4	A5	A6	A7	*	B1	B2	B3	B4	B5	B6	B7
Base	C	T	T	G	A	A	A	N	C	C	C	A	T	N	T

b) Sequence:

D	N	C	D	G	0	C	R	**GAP**	S	C	R	0.49
					0.03							0.19
					0.57			**L: 18-22**				0.02
					0.04							0.39
3.0			**4.44**						**3.0**			
									B2-B4?			

c) Sequence + Structure:

MINORWID	**MINORWID**	**MELTING**	**GAP**	**MELTING**	**GAP**	0.77	W	0.20	S	0.29
						0.89		0.15		0.83
L: 2	**L: 2-3**	**L: 6**	**L: 0-3**	**L: 6**	**L: 13**	0.56		0.80		0.20
R: 4.62/5.35	**R:4.62/5.35**	**R:72.78/82.26**		**R: 72.78/82.26**		0.55		0		0.52
81.2%	**75.0%**	**31.9%**		**31.9%**		**4.41**				
		B1-B5?		**B1-B5?**						

However, a total of 1323 hexamers fall within the range allowed by the 6 nt melting temperature constraint, so this motif component may be detecting a feature other than the –10 box.

The classification accuracy of the two models trained on *B. subtilis* upstream regions is shown in Table 6-10. As with the σ^{32} models, the number of false positive classifications was halved from the sequence-based model to the model using both sequence and structure (7.3% to 3.6%). The combined model also achieved perfect classification accuracy on the 22 positive training set sequences, one of which was not recognized by the sequence model. The best sequence and combined models are shown in Figure 6-4*b* and 6-4*c* respectively. The second motif component in Figure 6-4*b* contains three consecutive thymines, which are also found in the *B. subtilis* σ^{B} consensus. The weight matrix site immediately upstream of the TTT signal assigns a maximal score to guanine, which also agrees with the established consensus. The gap in the motif model is much shorter than the 12-15 nt gap in the promoter, suggesting that the downstream component was not detected by this model. However, recognition of the upstream conserved element and some adjacent residues was sufficient to yield ~90% classification accuracy on the training set. The combined model had two features that were highly restrictive: a stretch of 7-10 nt with a required helical Twist between 37.92° and 38.15°, and a major groove width requirement of exactly 12.15 Å over four or five positions. This major groove width can only be achieved by an unbroken A-tract (or T-tract), which is surprising because such a feature is not explicitly found in the promoter consensus.

Table 6-10: Classification accuracy obtained from two MGA models trained on 400 nt upstream regions containing *B. subtilis* 168 σ^B promoters. For both models, the percentage of positive cases correctly identified is shown, as well as the percentage of negative cases where patterns were incorrectly identified. The overall accuracy is the difference between these two values.

	Sequence Model	Sequence + Structure Model
True Positive	95.5%	100%
False Positive	7.3%	3.6%
Accuracy	88.2%	96.4%

Figure 6-4: The best scoring models trained on upstream sequences containing *B. subtilis* 168 σ^B promoters. The promoter consensus is shown in Figure 6-4*a*, with the variable length spacer region represented as an asterisk. Figure 6-4*b* shows a model trained with sequence components only, while Figure 6-4*c* shows the best model trained with sequence and structural components. The notation is described in Chapter 5. For sequence-based motif components, the string of IUPAC and/or weight matrix positions are shown in order, with the minimum score required for a match shown below the motif component. IUPAC letters and their corresponding bases are shown in Table 1-1; a base that matches the appropriate IUPAC letter is assigned a score of 1.0, while a mismatch is assigned a score of 0.0. A weight matrix position shows the score associated with adenine, cytosine, guanine, and thymine in four rows. Gaps are shown with their associated length (L:) in nucleotides. Structural motif components also have an associated length, as well as an acceptable range (R:) for the averaged parameter value (See appendix A for structural mapping rules). The bottom row of a structural component indicates what percentage of all possible *n*-mers of length L are within the acceptable range R. In Figures 6-4*b* and 6-4*c*, the sites that are consistent with the σ^B consensus are underlined, and the positions they match (if any) are indicated beneath the appropriate motif component.

a) Consensus:

Position	A1	A2	A3	A4	A5	A6	*	B1	B2	B3	B4	B5	B6
Base	G	T	T	T	W	W	N	G	G	G	W	A	W

b) Sequence:

Y	B	0	0	T	T	T	0	0	**GAP**	D	H	N	N	V
		0.29	0				0.65	0.62						
		0.49	1				0.10	0.47	**L: 5**					
		0.26	0.87				0.70	0.18						
1.70		**5.07**								**4.77**				
		A1-A6												

c) Sequence + Structure:

MINORWID	**MINORSIZE**	**MAJORDIST**	**TWIST**	**MINORDEP**	**MAJORWID**
L: 3	**L: 2**	**L: 2**	**L: 7-10**	**L: 4-7**	**L: 4-5**
R: 4.62/5.04	**R: 2.98/3.98**	**R: 3.38/3.79**	**R: 37.92/38.15**	**R: 8.96/9.11**	**R:12.15/12.15**
50.0%	**87.5%**	**56.25%**	**2.3%**	**83.9%**	**0.49%**
					?

Conclusions

Summary

The experiments performed in this chapter were intended to show the range of tasks MGA can perform on real sequence data. Several CAP binding site models were constructed with a variety of constraints, illustrating the different ways a regulatory site can be represented. Model searches against the set of upstream regions from a genome can be used to suggest other possible targets for regulatory sequences. The IHF models illustrated the role of structure in protein binding sites, with a model created from a single structural description yielding better classification accuracy than a model based on sequence alone. Finally, models were constructed from the extracted upstream regions of co-regulated genes to identify conserved patterns contained therein.

Classification accuracies were far better than random for all generated models, with between 80% and 100% of positive training set cases classified correctly, and between 85% and 98% negative training set cases correctly identified by most models. With the exception of the *E. coli* K12 σ^{32} promoter set, all data sets produced models with components that could clearly be tied to known consensus or structural features.

Building Models

Several key properties need to be considered when building an MGA model. While the system was able to represent the same binding site in several different ways, the goal of training should be clearly defined before training is started. As illustrated in all three sets of experiments performed above, model clarity is important if the goal is to understand the important features of a motif. Sequence features such as IUPAC strings and weight matrices are easy to understand, as they can be directly compared to the original sequences to determine where matches occurred. Since the majority of known consensus patterns are displayed as IUPAC strings or weight matrices, MGA models created with these constraints in place are easier to compare to known features. While structural representations can yield more accurate models, the features that are developed

to describe motifs are more difficult to understand. Since a given structural constraint can permit matches to many divergent sequences in addition to a consensus pattern, it can be hard to decide whether a structural motif component is meant to recognize a conserved region. Areas near a binding site that are unimportant to the function of that site may be assigned weak structural constraints by MGA, since a small training set is unlikely to cover the entire range of sequence possibilities. In contrast, conserved regions of a binding site often had very strong structural constraints.

Another important consideration is the choice of structural models to allow during MGA training. In the IHF training runs, the combined model yielded worse classification accuracy than the model constrained to sequence and flexibility, even though the combined model could have generated the motif as well. Though large numbers of Chromosomes and Evolvables allow MGA to explore many possible motif models, the heuristic method appears to have difficulty in choosing among a large set of available sequence and structural features. If a hypothesis about the role of certain structural features can be generated through other means, then the number of available structural rules can be limited to those suggested by the hypothesis.

The fitness function is another key determinant of the accuracy of the final model. The simple fitness function described at the beginning of this chapter allowed the generation of models with overall classification accuracies in excess of 90%. An alternate fitness function, intended to minimize the number of false positives, sacrificed the detection of important positive set members in favour of a reduced number of false positive classifications. This drive yielded models with much lower accuracy than those trained with the simpler fitness function.

Searching the Genome

Where tested, the generalization ability of the trained models was somewhat disappointing. All of the upstream regions in the test set had predicted CAP binding sites according to at least one of the four models trained with migration, but so were the majority of upstream regions not in the test set. While CAP regulates the expression of >300 genes in *E. coli*, it is implausible to suggest that it regulates >3000 genes by binding

in their upstream regions, especially when downstream operonic genes are considered. The majority of binding sites predicted by this model are therefore either false positive predictions, where CAP does not bind, or legitimate CAP binding sites that arose by chance or by coincidence and play no regulatory role.

In the combined poll model, 25 separately trained MGA Chromosomes were used to predict sites in the genome of *E. coli* K12. However, due in part to the fitness function used and in part to the lack of migration between Evolvables, none of these 25 motif models achieved the same level of classification accuracy as any of the models trained with migration. This 'panel of experts' performed poorly on the genome, and in most cases the combined predictions failed to distinguish upstream regions containing CAP binding sites from the rest of the genome. This approach may be worthwhile if more accurate models (such as the constrained models trained with the simpler fitness function) are used, which would require more training to yield better prediction accuracy.

Improving Model Accuracy

The choice of negative training set is vital to the development of meaningful motif models. While upstream regions tend to be less conserved than protein-coding sequences, they are not random, and the motif or motifs of interest are unlikely to be the only patterns contained in these sequences. For instance, when coding sequence was used as the negative training set for a set of upstream regions (not shown), the 'motifs' detected were invariably long tracts of adenines or thymines, which are more common in upstream regions than in coding sequences. While some of these may be involved in interactions with RNA polymerase (Aiyar *et al.*, 1998), others are likely not regulatory signals (Marchal *et al.*, 2003). The negative set sequences should contain the same biases as the positive training set, with the exception of the motifs of interest.

While shuffled sequence is a better approximation because it preserves the nucleotide composition, the shuffling process can still disrupt higher-order patterns that are not related to the motifs of interest, such as A-tracts, dinucleotides, or structural features common to most upstream regions (Pedersen *et al.*, 2000). The ideal negative set would consist of upstream regions known to lack binding sites for the protein of interest.

Such a set can be difficult to assemble: while published sequences known to contain regulatory sites are abundant, sequences known to *lack* binding sites for a specific protein are generally not presented. The absence of evidence is not sufficient to conclude that a protein cannot bind a given 400 nt upstream region, especially if, as suggested above, regulatory proteins can bind DNA sequences without assuming a regulatory role.

[7]

Discussion

Summary of Results

This thesis introduced two methods designed to detect conserved patterns in DNA. The Genetic Algorithm Neural Network (GANN) system identifies combinations of patterns that can yield good discrimination between different categories of DNA sequence, and uses a windowing system to reduce the amount of noise present in the final model. The Motif Genetic Algorithm (MGA) method also searches for combinations of patterns, but the MotifGene representation is better suited to finding adjacent patterns that combine to yield a single motif.

The GANN experiments described in Chapters 3 and 4 showed how success in pattern detection depends on the information conveyed by the encoded data. The representation used in Chapter 3 was not effective in allowing discrimination between upstream and non-upstream regions, which was evident from comparisons of the experimental and negative control runs. The advantages of a position-independent encoding were negated by the inability of the interval encoding scheme to represent coherent motifs. The simple indices used in Chapter 4, combined with the influence of the windowing system, allowed the detection of motifs with conserved positions relative to the start codon, such as the ribosome binding site in *Escherichia coli* K12 and the promoter consensus of *Sulfolobus solfataricus* P2. However, there was little evidence to suggest that rare motifs or motifs with variable positioning were detected.

The experiments reported in Chapters 5 and 6 were used to explore and understand the capabilities of MGA, and to show how different properties of MGA could be exploited to handle the different challenges of motif detection. The positive control experiments in Chapter 5 illustrated the ability of MGA to address the problems of degenerate motifs, variable motif spacing, multiple motifs and higher-order descriptions

of DNA. These principles were applied to generate the models seen in Chapter 6, which described real biological sequences exhibiting some or all of the problems listed above. MGA was able to identify known, conserved sequence patterns in CRP binding sites and in the promoters recognized by *E. coli* σ^{32} and *B. subtilis* σ^{B}. The inclusion of DNA structure encoding in model construction yielded final models that performed better on the training set than those based on sequence alone.

The primary goal of this thesis project was to develop motif detection methods that address four key challenges of biological pattern detection, namely i) motif degeneracy, ii) variable motif positioning within upstream regions, iii) different motif combinations in different upstream regions, and iv) the role of higher-order properties of DNA such as structure and flexibility. The core pattern detection functions in GANN are separate from the sequence encoding, and are suited to problem (iii) since GANN seeks combinations of patterns that define different categories of DNA sequence (represented by indices). The sequence encoding used in Chapter 4 was intended to address the other three issues through degenerate characters (i), different window sizes (ii), and structural indices (iv). The success of degenerate modeling was best illustrated by the indices that were used to model the ribosome binding site: many purine-rich indices close to the start codon were indicative of different RBS sequences, and spread through the population of OGA Chromosomes due to their high fitness. However, the windowing system was unable to model any motifs that had variable positioning within the upstream sequences, and there was little evidence that structure was a useful contributor to prediction accuracy.

Rather than relying on precalculated indices with defined window sizes, MGA optimizes the length, range and identity of motif components. It therefore possesses more flexibility to search through the training set sequences and to develop a highly specific, optimal representation. MGA is thus better suited to addressing problems i) and iv) above, with its ability to introduce slight modifications to its motif components. Since MGA can produce indices that are more specific than those used with GANN above, the need for a windowing system is reduced, since more specific indices will occur less frequently by chance in the target sequences.

MGA/GANN: Heuristic Methods

GANN and MGA are both heuristic methods that use iteration to find patterns that discriminate well between different types of sequence. While these methods will usually reach some optimal level of prediction accuracy, there is no guarantee that the best possible solution will be generated. The genetic algorithm that underlies both of these methods is used to help the systems 'escape' from local optima by preserving multiple sets of model parameters that can be recombined. The solution obtained with a GA approach is less likely to depend on the (random) starting parameters of the model than a gradient descent method such as backpropagation (Bishop, 1995), though the starting conditions of a GA will still influence the final model. However, the flexibility of GA has a serious computational cost, as every model in each generation of the population must be tested on the target data. Thus, the time required to evaluate a GANN or MGA Chromosome is multiplied by the number of such Chromosomes. Even with this added computational load, the models generated are not guaranteed to be optimal, as the population of GA Chromosomes may homogenize over time, thus eliminating the potential for new parameter sets that search a new region of the solution space. However, the use of heuristic methods is justified in these approaches. In the case of GANN, heuristics are necessary to train the neural networks, which were shown in Chapter 4 to yield better generalization than deterministic methods such as discriminant function analysis. While deterministic methods can be used to develop weight matrices (Stormo and Hartzell, 1989) based on sequence alone, the presence of multiple alternative representations (gaps, strings, weight matrices and structural encoding) creates the need for a heuristic method to test different combinations of these units.

Another advantage of the heuristic method in MGA is mentioned in Chapter 6 in relation to genomic predictions. While a single motif model may be too specific to the training sequences (thus yielding poor generalization across the rest of the genome), the use of multiple Evolvables without migration to train independent models can produce a 'panel of experts'. Even though the trained MGA Chromosomes may recognize the same motif, these 'experts' may focus on different parts of this motif. While polling this set of

models will yield more false positives, a sequence that matches many independent models may be more likely to contain a real example of the modeled motif.

Biological Context

Motifs as Instances of a General Model

An important objective of these experiments was to generate models of motifs and upstream regions that could be used to predict other such features in genomic DNA. The accuracy of these predictions depend on the ability of the model to detect motifs that were not in its training set. If the training motifs are a good sample of all possible sequences that can interact with the regulatory protein, then the trained model will likely have good generalization ability. In contrast, choosing a narrow, non-random sample of all possible binding sites for training will yield a model that only recognizes the training cases, and can make few useful predictions on new sequence data.

Since many regulatory proteins, including alternative σ factors, appear to regulate only a few transcriptional units (Rosen and Ron, 2002), even the entire set of binding sites from the genome will not represent the complete range of possible binding sites. This restriction was dramatically illustrated with the SELEX experiments by Horwitz and Loeb (1986) who showed that *E. coli* σ^{70} could bind strongly to many DNA sequences that were not in the set of characterized promoters. Thus, MGA or GANN models that are trained on a set of extracted genomic sequences will not reflect the entire range of binding possibilities. Notwithstanding this limitation inherent in all predictive methods, MGA and GANN were both shown to be able to detect conserved patterns in the training sequences, which yields useful information about the identity and position of regulatory sites.

False Positives

Experiments involving MGA and GANN often yielded classification accuracies >90% on the positive set of sequences known or thought to contain related patterns of

interest. However, all experiments yielded false positive predictions as well, and the 'pattern of interest' was often detected in 5% to 15% of the negative set sequences. From a pattern detection point of view, these false positives could simply reflect a sub-optimal model, either due to the heuristic training method finding a local optimum, or a data representation that is unable to model all training set motifs perfectly to the exclusion of everything else. The replicated GANN runs in Chapter 4 did not all achieve the exact same generalization accuracy, nor did MGA Evolvables trained concurrently without migration. As described in **Heuristic Methods** above, these models are influenced by the random elements chosen at the beginning of the run, and the random changes induced through mutation. The limitations of the data representations used for GANN in Chapters 3 and 4 were discussed in those chapters. While MGA can model relationships between sites in a limited way using weight matrices, this system lacks the extensive and explicit interaction modeling of artificial neural networks. If interactions between different bases are important, then MGA may not be able to represent these dependencies accurately.

Beyond the limitations of the detection methods lie questions about the biological significance of false positives. Is it reasonable to assume that binding sites for proteins are only found upstream of the genes they regulate, and nowhere else in the genome? Since binding sites can vary considerably from the consensus, some detection schemes require a motif to match the consensus sequence with no more than a specified number of mismatches (Li *et al.,* 2002; Keich and Pevzner, 2002). Increasing the number of permitted mismatches will increase the number of detected motifs in a genome, but not all of these 'motifs' may in fact be involved in regulation. In addition to binding sites in promoters, the model will likely detect protein binding sites that have arisen by chance, or through coincidental similarity to another feature that is under selective pressure. A motif detected within a coding region, for instance, is less likely to be involved in regulation due to its distance from an active promoter, but may still represent a valid binding site for the regulatory protein. Unless there is some selective disadvantage to the presence of regulatory sequences within a coding region resulting from interference with legitimate transcription, titration of rare regulatory proteins, or wasting of energy due to transcriptional activation within an ORF (if an entire functional promoter is present), then

there should be no strict barrier to the occurrence of regulatory sequences within coding regions.

"Yes or No" Predictions

By classifying sequences or sets of indices into two or more groups, GANN and MGA implicitly make "yes or no" predictions about whether a motif is present. This strategy is the simplest available approach, and is widely implemented in the literature. However, this type of prediction ignores the fact that the strength of a DNA-protein interaction is variable, depending on the sequence of the target motif. For instance, σ^{70} promoters from *E. coli* K12 are stronger if they are similar to the consensus sequence, and weak if several nucleotides differ from this ideal. Simple "on/off" models can ignore the relative strength of the regulatory sequences being modeled. Weight matrices are an exception to this rule, since they generate a floating-point number that may indicate the strength of the site (Ponomarenko *et al.*, 1999*a*); however, the correlation between weight matrix prediction and binding strength is unlikely to hold if the matrix has been trained on only a few motifs.

If binding affinity data are available, they can be used to generate correlative models. Ponomarenko *et al.* (1999*b*) performed such analyses with eukaryotic motifs, showing that the binding affinity between a regulatory protein and a given motif could sometimes be correlated with physicochemical properties of the motif such as helical twist and minor groove depth. In the absence of affinity data, an "on/off" model may be the best that can be constructed with any reliability, since the binding affinities are not available to build a correlation model. However, such models may be forced to miss many boundary cases of motifs with very low binding affinity, and must give equal weight to such cases if they are included in a training set with stronger binding sites.

Inferring Motif Function From a Model

In considering multiple alternate representations of DNA, GANN and MGA introduce a great deal of flexibility into the construction of models. In particular, the

inclusion of numerical conversions that represent DNA structure and flexibility allow the association of otherwise unrelated patterns in DNA with one another *via* different groupings of *n*-mers. This principle was best illustrated in the final experiment of Chapter 5: sequences with low nucleotide similarity can in fact have similar overall structural properties, and a model that recognizes this similarity can yield better discrimination between binding sites and other sequence.

Any number of these structural relationships can be defined for *n*-mers of any length, each providing a different way of ordering and grouping the DNA words, even to the point of associating randomly chosen values with each word to yield new categories. However, the abundance of categories raises three issues that must be considered:

1) The computational time and feasibility of searching through different combinations of these categories. Increasing the number of categories may eventually yield diminishing returns, as the search method may not be able to consider all of them during a training run.
2) Structural categories are not independent, and introducing a large number of categories will tend to muddle the motif models without increasing their accuracy. If three different representations are equally good for modeling a motif component, then the model will end up choosing one or more at random. This effect leads to models that are too complicated, and makes it harder to compare between trained models.
3) The difficulty of interpreting the motifs increases with the number of mapping rules available to the training algorithm. While a composite model that includes five different types of structural representation may be useful in detecting the training cases and searching the genome, understanding the role of each of its components in DNA-protein interactions can be very difficult.

In general, if the goal of an MGA or GANN run is to discover conserved patterns and use these patterns to scan the rest of the genome, then understanding the exact choice of structural features and functional role of structure in the motif is not vital. However, if an understanding of the motif itself is sought, then the features available must be

carefully selected before the experiment is performed. In Chapter 6, MGA runs constrained to sequence models only were performed prior to the construction of more 'free' models. The sequence-based models thus generated could be more easily compared to known motifs from the literature (Stormo and Hartzell 1989; Gross, 1996; Bewley *et al.*, 1998; Petersohn *et al.*, 1999) than could the structural models. This step was needed to demonstrate that MGA was detecting the correct motif. Similarly, if a hypothesis exists about the role of structure in a binding site prior to model construction, then the training run should only be allowed to use the mapping rules predicted to be useful by the hypothesis. This principle was illustrated in the IHF modeling runs in Chapter 6 where only the flexibility calculations from Olson *et al.* (1998) among all available rules were permitted in the model. This category was chosen because of the demonstrated role of deformation energy in IHF binding sites (Steffen *et al.*, 2002). As a further example, several studies have suggested a role for DNA flexibility in the spacer region between the –35 and –10 boxes of the σ^{70} promoter and near the eukaryotic TATA box (Nussinov, 1992; Ayers *et al.*, 1989). Thus, an MGA model to test this hypothesis should permit only flexibility and DNA sequence to be considered during training.

Reliability of the Data

If the underlying data used to construct a model are suspect, then the detection method trained on these data may be rewarded for detecting modified and possibly incorrect signals in the training set. MGA and GANN rely on three types of data to train their models: the raw DNA sequence, the annotated coding sequences within a genome, and the rules used to convert DNA sequence to structure.

Since the DNA sequence is the ultimate source of all three types of information, obtaining highly accurate DNA sequence is vital to the construction of meaningful models. DNA sequencing is susceptible to errors with rates dependent on several factors, including the positioning of the nucleotide relative to the start of the current sequencing reaction, the type of polymerase used for sequencing, and the G+C content of the template DNA sequence (Richterich 1998). Errors in coding sequences, particularly insertions or deletions, can often be detected by comparison with coding sequences from

related organisms (Médigue *et al.*, 1999). However, errors in upstream regions are much harder to detect, since there is no selective pressure to preserve the reading frame, and observed insertions or deletions will be less likely due to sequencing errors. Thus, while the error rate of DNA sequencing is quite low, with 1/10 000 nt a target error rate for some genome sequencing projects (Ewing *et al.*, 1998), this rate will be hardest to assess, and likely at its highest, in intergenic regions.

An accurate description of the predicted coding sequences is essential to the extraction of the correct upstream regions of genes or operons. Two annotation problems are most likely to yield incorrect definitions of coding versus intergenic sequence. The first challenge lies in the correct identification of short protein-coding sequences within the genome. Short ORFs are more likely to occur by chance than long ones, due to the likelihood of an in-frame stop codon occurring by chance. However, some short ORFs do code for proteins, and it is far too simplistic to assign a minimum length cutoff in determining whether an ORF codes for a protein. This approach was taken in the annotation of the *Aeropyrum pernix* genome, and led to massive overpredicting of coding sequences, with roughly 50% of all annotated genes thought to be spurious (Bocs *et al.*, 2002; Skovgaard *et al.*, 2001). Another property of short ORFs that makes them harder to characterize is their decreased likelihood of matching known gene products in databases such as GenBank. Since short sequences are more likely to match one another by chance, BLASTP may not assign a high significance level to valid matches between short sequences. If such a match is thus rejected, a coding sequence is missed.

Even if a coding sequence is found, the correct start codon may not be selected. If several ATG or GTG codons are present at the 5′ end of a coding sequence, then it may be difficult to choose between them. The annotation can be assisted with 'extrinsic' comparisons with orthologous genes from other genomes, but these comparisons can be of limited use since the 5′ end of a coding sequence is often only weakly conserved (Hayes and Borodovsky, 1998). While an incorrectly annotated start codon may not prohibit the study of the protein product of a gene, it will have a serious impact on the definition of the upstream region. In such a case, the 'upstream region' as defined may in fact contain coding sequence, or it may miss intergenic sequence that is close to the correct start codon.

Some of the issues concerning the use of structural measures of DNA were addressed in the Introduction, in particular the limitations of crystallographic data. While these data have been previously analyzed to yield mean values of static parameters such as helical twist (el Hassan and Calladine, 1996; Gorin *et al.*, 1995), others have examined the range and distribution of values in an attempt to determine conformational mobility (Olson *et al.*, 1998; Ponomarenko *et al.*, 1999*b*). With this approach, the relative effect of different crystal environments on different base pair steps yields important information about the mobility of these steps, since some steps are more strongly influenced by packing forces or protein binding than others. The parameters that describe flexibility and conformational variability may therefore be more reliable than those that describe static helical parameters. Since protein-DNA interactions can be dependent on factors such as supercoiling (Opel *et al.*, 2001), the conformational mobility may again be a more useful way to represent the DNA molecule.

Further Directions for MGA/GANN

Applications of GANN

The GANN experiments described in this thesis relied on a series of simple sequence and structural measures to represent the composition of upstream regions. These experiments showed that GANN could detect and classify patterns in encoded DNA, but they revealed the limitation of using these 'naïve' indices. GANN was presented with a challenge that it could not overcome: not only to determine which combinations of protein binding sites combine to yield a functional regulatory sequence, but also which combinations of basic DNA patterns define the myriad binding sites present in a genome such as that of *E. coli* K12. This approach was designed to address the problem of characterizing regulatory sequences without prior experimental knowledge, but GANN was unable to reconstruct most binding sites from simple indices, and so it failed to identify co-regulated sets of genes. The heavy reliance on a few related indices (especially those connected to the ribosome binding site in *E. coli* K12) observed

in Chapter 4 suggests that the importance of interactions between sites was not captured in that experiment.

One possible solution to this problem is a division of labour: continue using GANN to determine which combinations of features are valid, but present GANN with a more meaningful set of indices. Ideally, every index would correspond to the presence or absence of a predicted binding site for a different regulatory protein, and GANN would make a positive or a negative prediction based on the identity and spacing of these different elements. The task of defining the binding sites would then fall to a method designed specifically to find motifs in DNA. Weight matrix representations are available for many characterized protein binding sites in *E. coli* K12 (Robison *et al.*, 1998). Li *et al.* (2002) presented a method for extracting weight matrices from the upstream regions of a genome which relied on identification of palindromes and near-palindromes. Many of the extracted weight matrices represented known binding sites, and others revealed candidate binding sites. However, since this approach depends on palindrome recognition, it does not detect most promoter sequences, and would miss sites where structure plays an important role such as IHF. Another option would be to use a detection method such as MGA to identify candidate motifs (see *Applications of MGA* below). Since GANN can accept data from multiple sources, a combination of these methods could be used to generate indices.

A key advantage of using a set of complete binding site representations would be the reduction in the number of indices presented to GANN. This reduction would diminish the amount of computational time needed for a training run of GANN, and would also limit the number of highly correlated indices obtained when basic sequence and structural characteristics are concerned. Also, the need for strict window separation could be relaxed, since the algorithm would focus more on the presence or absence of a well-defined feature, rather than trying to identify 'spikes' in the frequency of certain nucleotides or short structural features that are present everywhere. Instead of windows that identify the absolute position of a feature with respect to the start codon, different indices could consider the positioning of different features relative to one another. The importance of this type of measurement is shown in the position-specific role of CRP binding sites, which activate transcription when present at certain positions upstream of

the promoter, but repress transcription if they are downstream of the promoter (Botsford and Harman, 1992).

Applications of MGA

As demonstrated in Chapter 6, MGA is able to detect weakly conserved patterns with variable positioning in DNA. This ability was exploited to extract conserved motifs from upstream regions of genes known to be regulated by common sigma factors. These experiments could be repeated for sets of genes under the control of other sigma factors or of regulatory proteins to yield models for conserved motifs bound by these proteins. The best source of information for co-regulated genes is the Regulon Database (Salgado *et al.*, 2001). Over a hundred regulatory proteins of *E. coli* K12 are catalogued in this database, along with the genes they are known to regulate, and characterized binding sites. Models can thus be constructed from these data in two ways: either by considering the binding sites as they are presented in this (or any other) database, or by extracting suitably long upstream sequences from co-regulated genes.

Methods that can use extracted genomic sequence data may be particularly useful where microarray data are concerned. If the genome sequence of an organism is known, then microarray experiments that show co-expression of genes can indicate which sets of upstream regions should be searched for common binding sites. This principle has been applied to detect conserved regulatory motifs in yeast (Jensen and Knudsen, 2000) and other model organisms. However, a set of genes that are co-expressed in response to a stimulus need not all be directly controlled by the same regulatory protein: an important principle of gene expression is the regulatory cascade, where a regulatory protein controls the expression of one or more other regulatory proteins, thus implicating several regulatory proteins in a single response. This principle is illustrated in the cascade of sigma factors during sporulation in *Bacillus subtilis*, where coordinated sigma factor activity is essential to properly timed expression (Moran, 1993). If several regulatory proteins are used by a cell to trigger a regulatory response, the genes whose expression level changes during that response will have distinct sets of motifs in their upstream

regions. In such a case, the subsampling ability of MGA might allow the detection of these multiple motifs.

Since the motifs defined in MGA Chromosomes can be searched throughout a target genome, previously unknown binding sites can be detected in the upstream regions of genes not in the training set. As seen in Chapter 6, such predictions can be valid binding sites that contribute to the regulation of a gene, but they may also be found upstream of a gene that is unlikely to be under the control of the regulatory protein in question. MGA may thus 'overpredict' the number of regulatory binding sites in the genome. As discussed in *False Positives* above, these sites may be legitimate binding sites that arose by chance or coincidence, and do not contribute to regulation due to improper positioning relative to the promoter. If real binding sites are responsible for overprediction, then accurate predictions would depend on the interactions between different binding sites and the promoter. Thus, the GANN model described above with explicit descriptions of spacing between different motifs could be a useful approach to this problem. However, if the genomic predictions are mostly 'false positives' that cannot interact with a regulatory protein, then a better approach would be to reduce the number of false positives during training. This reduction could be attempted either by increasing the number of training rounds or the number of MGA Chromosomes, or by changing the fitness function to strongly favour MGA Chromosomes with very low false positive prediction rates.

Conclusion – Whole-Genome Motif Detection

Motif prediction and modeling are key components in understanding the control of gene expression within a cell. While transcription is not the only regulated step in gene expression, it is a key step and the primary target for regulation in both prokaryotes and eukaryotes. The number of transcriptional regulatory proteins in *E. coli* K12 is thought to be as high as 400 (Perez-Rueda and Collado-Vides, 2000), which illustrates the importance and the complexity of transcriptional regulation in this genome. The number of characterized transcription factors in other genomes is growing as well (Wingender *et*

al., 1996), suggesting that *E. coli* K12 is not exceptional in terms of its regulatory complexity.

If a single genome contains over 400 transcription factors, how can their binding sites all be detected and characterized? Many common regulatory proteins and sigma factors have been well characterized, to the point of knowing dozens or even hundreds of target sites for *E. coli* σ^{70} (Lisser and Margalit, 1994) and CAP (Salgado *et al.*, 2001), and eukaryotic proteins such as human TBP (Perier *et al.*, 2000). Crystal complexes of promoter/RNA polymerase complexes have yielded detailed information about these regulatory interactions. As demonstrated in the last two chapters, even if only a few examples of a transcription factor binding site are available, weight matrices and other representations can be constructed (Robison *et al.*, 1998) to generate a model of the binding site.

If the binding sites themselves are not known, then upstream regions of co-regulated genes can be searched for common sequence and structural patterns. In addition to the likelihood described above that different upstream regions may have altogether different motifs due to regulatory cascades, this type of search may also detect conserved patterns unrelated to the motifs of interest. The compositional difference between coding regions (50% G+C in *E. coli* K12) and intergenic regions (40% G+C) makes the protein-coding sequences wholly unsuitable as negative set members. Intergenic regions are compositionally different from coding regions (Marchal *et al.*, 2003), and these may be modeled if the negative training set is not carefully designed to include these features as well. Shuffled sequences, as applied in the MGA experiments in Chapters 5 and 6, may break up these long runs, thus underrepresenting them in the negative training set.

Some transcription factors may regulate only one or a few genes within a genome. The motifs bound by these proteins will be the most difficult to detect and model, since there is not enough information within a single genome to construct an adequate model. This limitation has prompted the development of phylogenetic footprinting. This method compares the upstream regions of orthologous genes from related strains or species of an organism. Since noncoding regions are on average less conserved than coding sequences (Jordan, 2002), the upstream regions of related strains will tend to have many nucleotide substitutions, except where sequence must be conserved as in promoters and other

binding sites. This approach was used by Mironov *et al.* (1999) to develop more refined models of the sites recognized by several conserved transcription factors from *E. coli* K12, *Haemophilus influenzae* Rd KW20 and several other related proteobacterial genomes. MGA could be used for phylogenetic footprinting as well to yield an expanded set of motif models.

At present, there is no shortage of sequence and expression data which can be used to generate motif models. MGA and GANN are tools suited to the analysis of these data, which can be used in concert with wet-lab analyses to propose and to validate hypotheses about co-regulated genes and important motif properties. If the parameters and set definitions for these algorithms are understood and selected carefully, then these methods can be powerful tools in assessing the transcriptional landscape of genomes.

[A]

Structural Mapping Rules

Table A-1: Mapping rules used to convert DNA dinucleotides to structural representations. The name of the rule is shown in the first column, followed by the ten conversions for each pair of reverse complementary dinucleotides. The reference for each mapping rule is shown in the final column.

Name	Dinucleotides										Reference
	AA/TT	AC/GT	AG/CT	AT	CA/TG	CC/GG	CG	GA/TC	GC	TA	
BEND	3.07	2.97	2.31	2.6	3.58	2.16	2.81	2.51	3.06	6.74	Karas *et al.*, 1996
BRUNAKPROP	-18.66	-13.10	-14.00	-15.01	-9.45	-8.11	-10.03	-13.48	-11.08	-11.85	Baldi *et al.*, 1998
BRUNAKSTACK	-5.37	-10.51	-6.78	-6.57	-6.57	-8.26	-9.69	-9.81	-14.59	-3.82	Baldi *et al.*, 1998
CLASH	0.64	0.95	2.53	1.68	0.80	1.78	2.42	0.03	0.22	0.00	Gorin *et al.*, 1995
ENTROPY	-21.9	-25.5	-16.4	-15.2	-21.0	-28.4	-29.0	-23.5	-26.4	-18.4	Sugimoto *et al.*, 1996
FLEXOLSON	2.9	2.3	2.1	1.6	9.8	6.1	12.1	4.5	4.0	6.3	Olson *et al.*, 1998
LISSERBTOA	-0.96	4.52	0.79	6.82	5.10	2.26	-10.79	3.18	8.28	0.42	Lisser and Margalit, 1994
LISSERBTOZ	16.3	5.4	14.2	10.0	19.2	10.0	16.7	10.5	2.9	24.7	Lisser and Margalit, 1994
LISSERFLEX	7.6	14.6	8.2	25.0	10.9	7.2	8.9	8.8	11.1	12.5	Lisser and Margalit, 1994
MAJORDEP	9.12	9.41	8.96	8.96	8.67	8.45	8.81	8.76	8.67	9.60	Karas *et al.*, 1996
MAJORDIST	3.38	3.03	3.36	3.02	3.79	3.38	3.77	3.40	3.04	3.81	Gorin *et al.*, 1995
MAJORSIZE	3.98	3.98	4.70	4.70	3.98	3.98	4.70	3.26	3.26	3.26	Gorin *et al.*, 1995
MAJORWID	12.15	12.37	13.51	12.87	13.58	15.49	14.42	13.93	14.55	12.32	Karas *et al.*, 1996
MELTING	54.50	97.73	58.42	57.02	54.71	85.97	72.55	86.44	136.12	36.73	Gotoh and Tagashira, 1981
MINORDEP	9.03	8.79	8.98	8.91	9.09	8.99	9.06	9.11	8.98	9.00	Karas *et al.*, 1996
MINORDIST	2.94	4.22	2.79	4.20	3.09	2.80	3.21	2.95	4.24	2.97	Gorin *et al.*, 1995
MINORSIZE	2.98	3.26	3.98	3.26	3.70	3.98	4.70	2.98	3.26	2.70	Gorin *et al.*, 1995
MINORWID	5.30	6.04	5.19	5.31	4.79	4.62	5.16	4.71	4.74	6.40	Karas *et al.*, 1996
NUCLEOSOME	18.4	10.2	14.5	7.2	15.7	10.2	1.1	11.3	5.2	6.2	Satchwell and Travers, 1989
ROLL	0.3	0.5	4.5	-0.8	0.5	6.0	-6.2	-1.3	-6.2	2.8	Gorin *et al.*, 1995
SLIDE	-0.1	-0.2	0.4	-0.4	1.6	0.8	0.7	0.0	0.4	0.9	Gorin *et al.*, 1995
TWIST	38.90	31.12	32.15	33.81	41.41	34.96	32.91	41.31	38.50	33.28	Gorin *et al.*, 1995

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.H., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Abril, J.F., Agbayani, A., An, H.J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N.V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacleb, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D., Scheeler, F., Shen, H., Shue, B.C., Siden-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., WoodageT, Worley, K.C., Wu, D., Yang, S., Yao, Q.A., Ye, J., Yeh, R.F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X.H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A., Myers, E.W., Rubin, G.M., and Venter, J.C. (2000) The genome sequence of Drosophila melanogaster. *Science* **287**: 2185-2195.

Aiyar, S.E., Gourse, R.L., and Ross, W. (1998) Upstream A-tracts increase bacterial promoter activity through interactions with the RNA polymerase alpha subunit. *Proc Natl Acad Sci U S A* **95**: 14652-14657.

Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S., and Ishihama, A. (1999*a*) Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid. *J Bacteriol* **181**: 6361-6370.

Ali Azam, T.A., and Ishihama, A. (1999*b*) Twelve species of the nucleoid-associated protein from Escherichia coli. Sequence recognition specificity and DNA binding affinity. *J Biol Chem* **274**: 33105-33113.

Altschul, S.F., and Koonin, E.V. (1998) Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem Sci* **23**: 444-447.

Amati, B., and Land, H. (1994) Myc-Max-Mad: a transcription factor network controlling cell cycle progression, differentiation and death. *Curr Opin Genet Dev* **4**: 102-108.

Amati, B., Frank, S.R., Donjerkovic, D., and Taubert, S. (2001) Function of the c-Myc oncoprotein in chromatin remodeling and transcription. *Biochim Biophys Acta* **1471**: M135-145.

Arguelles, J.C. (2000) Physiological roles of trehalose in bacteria and yeasts: a comparative analysis. *Arch Microbiol* **174**: 217-224.

Auble, D.T., and deHaseth, P.L. (1988) Promoter recognition by Escherichia coli RNA polymerase. Influence of DNA structure in the spacer separating the -10 and -35 regions. *J Mol Biol* **202**: 471-482.

Ayers, D.G., Auble, D.T., and deHaseth, P.L. (1989) Promoter recognition by Escherichia coli RNA polymerase. Role of the spacer DNA in functional complex formation. *J Mol Biol* **207**: 749-756.

Back, T. (2000) Introduction to evolutionary algorithms. In *Evolutionary Computation.* Vol. 1. Back, T., Fogel, D. B., Michaelewicz, Z. (ed): Institute of Physics Press.

Bailey, T.L., and Elkan, C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* **3**: 21-29.

Bajic, V.B., Seah, S.H., Chong, A., Zhang, G., Koh, J.L., and Brusic, V. (2002) Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. *Bioinformatics* **18**: 198-199.

Baldi, P., Chauvin, Y., Brunak, S., Gorodkin, J., and Pedersen, A.G. (1998) Computational applications of DNA structural scales. *Proc Int Conf Intell Syst Mol Biol* **6**: 35-42.

Baldi, P., and Brunak, S. (2001) Bioinformatics: The Machine Learning Approach. Cambridge, Massachusetts: MIT Press.

Baliga, N.S., and Dassarma, S. (2000) Saturation mutagenesis of the haloarchaeal bop gene promoter: identification of DNA supercoiling sensitivity sites and absence of TFB recognition element and UAS enhancer activity. *Mol Microbiol* **36**: 1175-

1183.

Baliga, N.S., Kennedy, S.P., Ng, W.V., Hood, L., and DasSarma, S. (2001) Genomic and genetic dissection of an archaeal regulon. *Proc Natl Acad Sci U S A* **98**: 2521-2525.

Baliga, N.S. (2001) Promoter analysis by saturation mutagenesis. *Biol Proced Online* **3**: 64-69.

Barber, A.M., and Zhurkin, V.B. (1990) CAP binding sites reveal pyrimidine-purine pattern characteristic of DNA bending. *J Biomol Struct Dyn* **8**: 213-232.

Bauer, W.R., Hayes, J.J., White, J.H., and Wolffe, A.P. (1994) Nucleosome structural changes due to acetylation. *J Mol Biol* **236**: 685-690.

Bautz, E.K., and Bautz, F.A. (1970) Initiation of RNA synthesis: the function of sigma in the binding of RNA polymerase to promoter sites. *Nature* **226**: 1219-1222.

Baylin, S.B., Esteller, M., Rountree, M.R., Bachman, K.E., Schuebel, K., and Herman, J.G. (2001) Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum Mol Genet* **10**: 687-692.

Beckett, D. (2001) Regulated assembly of transcription factors and control of transcription initiation. *J Mol Biol* **314**: 335-352.

Beckwith, J. (1996) The operon: an historical account. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. Vol. 1. Neidhardt, F.C. (ed). Washington, D.C.: American Society for Microbiology, pp. 1227-1231.

Bell, S.D., Jaxel, C., Nadal, M., Kosa, P.F., and Jackson, S.P. (1998) Temperature, template topology, and factor requirements of archaeal transcription. *Proc Natl Acad Sci U S A* **95**: 15218-15222.

Bell, S.D., and Jackson, S.P. (2000) Mechanism of autoregulation by an archaeal transcriptional repressor. *J Biol Chem* **275**: 31624-31629.

Bell, S.D., and Jackson, S.P. (2001) Mechanism and regulation of transcription in archaea. *Curr Opin Microbiol* **4**: 208-213.

Bell, S.D., Botting, C.H., Wardleworth, B.N., Jackson, S.P., and White, M.F. (2002) The interaction of Alba, a conserved archaeal chromatin protein, with Sir2 and its regulation by acetylation. *Science* **296**: 148-151.

Benson, A.K., and Haldenwang, W.G. (1993) Bacillus subtilis sigma B is regulated by a binding protein (RsbW) that blocks its association with core RNA polymerase. *Proc Natl Acad Sci U S A* **90**: 2330-2334.

Berg, O.G., and von Hippel, P.H. (1988) Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *J Mol Biol* **200**: 709-723.

Berman, H.M., Westbrook, J., Feng, Z., Iype, L., Schneider, B., and Zardecki, C. (2003) The nucleic acid database. *Methods Biochem Anal* **44**: 199-216.

Besemer, J., Lomsadze, A., and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**: 2607-2618.

Bewley, C.A., Gronenborn, A.M., and Clore, G.M. (1998) Minor groove-binding architectural proteins: structure, function, and DNA recognition. *Annu Rev Biophys Biomol Struct* **27**: 105-131.

Bisant, D., and Maizel, J. (1995) Identification of ribosome binding sites in Escherichia coli using neural network models. *Nucleic Acids Res* **23**: 1632-1639.

Bishop, C.M. (1995) Neural Networks for Pattern Recognition. New York: Oxford University Press.

Black, A.R., Black, J.D., and Azizkhan-Clifford, J. (2001) Sp1 and kruppel-like factor family of transcription factors in cell growth regulation and cancer. *J Cell Physiol* **188**: 143-160.

Blackwood, E.M., and Kadonaga, J.T. (1998) Going the distance: a current view of enhancer action. *Science* **281**: 61-63.

Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., and Shao, Y. (1997) The complete genome sequence of Escherichia coli K-12. *Science* **277**: 1453-1474.

Bocs, S., Danchin, A., and Medigue, C. (2002) Re-annotation of genome microbial CoDing-Sequences: finding new genes and inaccurately annotated genes. *BMC Bioinformatics* **3**: 5.

Boos, W., and Shuman, H. (1998) Maltose/maltodextrin system of Escherichia coli: transport, metabolism, and regulation. *Microbiol Mol Biol Rev* **62**: 204-229.

Bose, N.K., Liang, P. (1996) Neural Network Fundamentals with Graphs, Algorithms, and Applications. New York: McGraw-Hill.

Botsford, J.L., and Harman, J.G. (1992) Cyclic AMP in prokaryotes. *Microbiol Rev* **56**:

100-122.

Brennan, R.G., and Matthews, B.W. (1989) The helix-turn-helix DNA binding motif. *J Biol Chem* **264**: 1903-1906.

Britton, R.A., Eichenberger, P., Gonzalez-Pastor, J.E., Fawcett, P., Monson, R., Losick, R., and Grossman, A.D. (2002) Genome-wide analysis of the stationary-phase sigma factor (sigma-H) regulon of Bacillus subtilis. *J Bacteriol* **184**: 4881-4890.

Bucher, P., and Trifonov, E.N. (1986) Compilation and analysis of eukaryotic POL II promoter sequences. *Nucleic Acids Res* **14**: 10009-10026.

Burley, S.K., and Kamada, K. (2002) Transcription factor complexes. *Curr Opin Struct Biol* **12**: 225-230.

Busby, S., and Ebright, R.H. (1999) Transcription activation by catabolite activator protein (CAP). *J Mol Biol* **293**: 199-213.

Bussemaker, H.J., Li, H., and Siggia, E.D. (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A* **97**: 10096-10100.

Bussemaker, H.J., Li, H., and Siggia, E.D. (2001) Regulatory element detection using correlation with expression. *Nat Genet* **27**: 167-171.

Bykowski, T., and Sirko, A. (1998) Selected phenotypes of ihf mutants of Escherichia coli. *Biochimie* **80**: 987-1001.

Cai, Y.D., Liu, X.J., and Chou, K.C. (2001) Artificial neural network model for predicting membrane protein types. *J Biomol Struct Dyn* **18**: 607-610.

Cao, Q., and Richter, J.D. (2002) Dissolution of the maskin-eIF4E complex by cytoplasmic polyadenylation and poly(A)-binding protein controls cyclin B1 mRNA translation and oocyte maturation. *Embo J* **21**: 3852-3862.

Cardon, L.R., and Stormo, G.D. (1992) Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J Mol Biol* **223**: 159-170.

Carpenter, G.A., Grossberg, S. (1988) The art of adaptive pattern recognition by a self-organizing neural network. *IEEE Computer* **21**: 77-88.

Cases, I., and de Lorenzo, V. (1998) Expression systems and physiological control of promoter activity in bacteria. *Curr Opin Microbiol* **1**: 303-310.

Cashel, M., Gentry, D.R., Hernandez, V.J., Vinella, D. (1996) The Stringent Response. In

Escherichia coli and Salmonella: Cellular and Molecular Biology. Vol. 1. Neidhardt, F.C. (ed). Washington, D.C.: American Society for Microbiology, pp. 1458-1496.

Chatterji, D., and Ojha, A.K. (2001) Revisiting the stringent response, ppGpp and starvation signaling. *Curr Opin Microbiol* **4**: 160-165.

Chen, Q.K., Hertz, G.Z., and Stormo, G.D. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput Appl Biosci* **11**: 563-566.

Cheng, S.W., Davies, K.P., Yung, E., Beltran, R.J., Yu, J., and Kalpana, G.V. (1999) c-MYC interacts with INI1/hSNF5 and requires the SWI/SNF complex for transactivation function. *Nat Genet* **22**: 102-105.

Chuprina, V.P., Rullmann, J.A., Lamerichs, R.M., van Boom, J.H., Boelens, R., and Kaptein, R. (1993) Structure of the complex of lac repressor headpiece and an 11 base-pair half-operator determined by nuclear magnetic resonance spectroscopy and restrained molecular dynamics. *J Mol Biol* **234**: 446-462.

Claassen, G.F., and Hann, S.R. (1999) Myc-mediated transformation: the repression connection. *Oncogene* **18**: 2925-2933.

Claverie, J.M. (1994) Some useful statistical properties of position-weight matrices. *Comput Chem* **18**: 287-294.

Claverie-Martin, F., and Magasanik, B. (1991) Role of integration host factor in the regulation of the glnHp2 promoter of Escherichia coli. *Proc Natl Acad Sci U S A* **88**: 1631-1635.

Corbi, N., Perez, M., Maione, R., and Passananti, C. (1997) Synthesis of a new zinc finger peptide; comparison of its 'code' deduced and 'CASTing' derived binding sites. *FEBS Lett* **417**: 71-74.

Cramer, P., Bushnell, D.A., and Kornberg, R.D. (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292**: 1863-1876.

Crowley, E.M., Roeder, K., and Bina, M. (1997) A statistical model for locating regulatory regions in genomic DNA. *J Mol Biol* **268**: 8-14.

Crowley, E.M. (2001) A Bayesian method for finding regulatory segments in DNA. *Biopolymers* **58**: 165-174.

Dahlke, I., and Thomm, M. (2002) A Pyrococcus homolog of the leucine-responsive regulatory protein, LrpA, inhibits transcription by abrogating RNA polymerase recruitment. *Nucleic Acids Res* **30**: 701-710.

Darst, S.A. (2001) Bacterial RNA polymerase. *Curr Opin Struct Biol* **11**: 155-162.

Daube, S.S., and von Hippel, P.H. (1999) Interactions of Escherichia coli sigma(70) within the transcription elongation complex. *Proc Natl Acad Sci U S A* **96**: 8390-8395.

Davis, L. (1987) Genetic Algorithms and Simulated Annealing. Los Altos, California: Morgan Kaufmann.

de Crombrugghe, B., Busby, S., and Buc, H. (1984) Cyclic AMP receptor protein: role in transcription activation. *Science* **224**: 831-838.

de Moor, C.H., and Richter, J.D. (1999) Cytoplasmic polyadenylation elements mediate masking and unmasking of cyclin B1 mRNA. *Embo J* **18**: 2294-2303.

de Ruijter, A.J., van Gennip, A.H., Caron, H.N., Kemp, S., and van Kuilenburg, A.B. (2003) Histone deacetylases (HDACs): characterization of the classical HDAC family. *Biochem J* **370**: 737-749.

Demeler, B., and Zhou, G.W. (1991) Neural network optimization for E. coli promoter prediction. *Nucleic Acids Res* **19**: 1593-1599.

Derreumaux, S., and Fermandjian, S. (2000) Bending and adaptability to proteins of the cAMP DNA-responsive element: molecular dynamics contrasted with NMR. *Biophys J* **79**: 656-669.

Dhavan, G.M., Crothers, D.M., Chance, M.R., and Brenowitz, M. (2002) Concerted binding and bending of DNA by Escherichia coli integration host factor. *J Mol Biol* **315**: 1027-1037.

Dickerson, R.E., Drew, H.R., Conner, B.N., Wing, R.M., Fratini, A.V., and Kopka, M.L. (1982) The anatomy of A-, B-, and Z-DNA. *Science* **216**: 475-485.

Dickerson, R.E. (1989) Definitions and nomenclature of nucleic acid structure components. *Nucleic Acids Res* **17**: 1797-1803.

Dickerson, R.E., Goodsell, D.S., and Neidle, S. (1994) "...the tyranny of the lattice..." *Proc Natl Acad Sci U S A* **91**: 3579-3583.

DiGabriele, A.D., Sanderson, M.R., and Steitz, T.A. (1989) Crystal lattice packing is important in determining the bend of a DNA dodecamer containing an adenine tract. *Proc Natl Acad Sci U S A* **86**: 1816-1820.

Dombroski, A.J., Walter, W.A., and Gross, C.A. (1993) Amino-terminal amino acids modulate sigma-factor DNA-binding activity. *Genes Dev* **7**: 2446-2455.

Dorman, C.J., Hinton, J.C., and Free, A. (1999) Domain organization and oligomerization among H-NS-like nucleoid-associated proteins in bacteria. *Trends Microbiol* **7**: 124-128.

Dvir, A., Conaway, J.W., and Conaway, R.C. (2001) Mechanism of transcription initiation and promoter escape by RNA polymerase II. *Curr Opin Genet Dev* **11**: 209-214.

Dvir, A. (2002) Promoter escape by RNA polymerase II. *Biochim Biophys Acta* **1577**: 208-223.

Dworkin, J., Ninfa, A.J., and Model, P. (1998) A protein-induced DNA bend increases the specificity of a prokaryotic enhancer-binding protein. *Genes Dev* **12**: 894-900.

Ebright, R.H., Ebright, Y.W., and Gunasekera, A. (1989) Consensus DNA site for the Escherichia coli catabolite gene activator protein (CAP): CAP exhibits a 450-fold higher affinity for the consensus DNA site than for the E. coli lac DNA site. *Nucleic Acids Res* **17**: 10295-10305.

Efimov, A.V. (1999) Complementary packing of alpha-helices in proteins. *FEBS Lett* **463**: 3-6.

Eisen, A., and Lucchesi, J.C. (1998) Unraveling the role of helicases in transcription. *Bioessays* **20**: 634-641.

el Hassan, M.A., and Calladine, C.R. (1996) Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J Mol Biol* **259**: 95-103.

el Hassan, M.A., and Calladine, C.R. (1998) Two distinct modes of protein-induced bending in DNA. *J Mol Biol* **282**: 331-343.

Ellenberger, T., and Landy, A. (1997) A good turn for DNA: the structure of integration host factor bound to DNA. *Structure* **5**: 153-157.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**: 175-185.

Flashner, Y., and Gralla, J.D. (1988) DNA dynamic flexibility and protein recognition: differential stimulation by bacterial histone-like protein HU. *Cell* **54**: 713-721.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., and et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**: 496-512.

Foury, F., Roganti, T., Lecrenier, N., and Purnelle, B. (1998) The complete sequence of the mitochondrial genome of Saccharomyces cerevisiae. *FEBS Lett* **440**: 325-331.

Frech, K., Danescu-Mayer, J., and Werner, T. (1997) A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J Mol Biol* **270**: 674-687.

Fuglsang, A., and Engberg, J. (2003) Non-randomness in Shine-Dalgarno regions: links to gene characteristics. *Biochem Biophys Res Commun* **302**: 296-301.

Gajiwala, K.S., and Burley, S.K. (2000) Winged helix proteins. *Curr Opin Struct Biol* **10**: 110-116.

Gallegos, M.T., Schleif, R., Bairoch, A., Hofmann, K., and Ramos, J.L. (1997) Arac/XylS family of transcriptional regulators. *Microbiol Mol Biol Rev* **61**: 393-410.

Gavathiotis, E., Sharman, G.J., and Searle, M.S. (2000) Sequence-dependent variation in DNA minor groove width dictates orientational preference of Hoechst 33258 in A-tract recognition: solution NMR structure of the 2:1 complex with d(CTTTTGCAAAAG)(2). *Nucleic Acids Res* **28**: 728-735.

Gehring, W.J. (1987) Homeo boxes in the study of development. *Science* **236**: 1245-1252.

Gelfand, M.S. (1999) Recognition of regulatory sites by genomic comparison. *Res Microbiol* **150**: 755-771.

Gelfand, M.S., Koonin, E.V., and Mironov, A.A. (2000) Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res* **28**: 695-705.

Ghosh, D. (1992) TFD: the transcription factors database. *Nucleic Acids Res* **20 Suppl**: 2091-2093.

Gingras, A.C., Raught, B., and Sonenberg, N. (1999) eIF4 initiation factors: effectors of mRNA recruitment to ribosomes and regulators of translation. *Annu Rev Biochem* **68**: 913-963.

Glover, J.N., and Harrison, S.C. (1995) Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature* **373**: 257-261.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. (1996) Life with 6000 genes. *Science* **274**: 546, 563-547.

Gorin, A.A., Zhurkin, V.B., and Olson, W.K. (1995) B-DNA twisting correlates with base-pair morphology. *J Mol Biol* **247**: 34-48.

Gotoh, O., and Tagashira, Y. (1981) Locations of frequently opening regions on natural DNAs and their relation to functional loci. *Biopolymers* **20**: 1043-1058.

Granjeon, E., and Tarroux, P. (1995) Detection of compositional constraints in nucleic acid sequences using neural networks. *Comput Appl Biosci* **11**: 29-37.

Griffith, K.L., and Wolf, R.E., Jr. (2002) A comprehensive alanine scanning mutagenesis of the Escherichia coli transcriptional activator SoxS: identifying amino acids important for DNA binding and transcription activation. *J Mol Biol* **322**: 237-257.

Gross, C.A. (1996) Function and regulation of the heat shock proteins. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. Vol. 1. Neidhardt, F.C. (ed). Washington, D.C.: American Society for Microbiology, pp. 1382-1399.

Grove, A., Galeone, A., Yu, E., Mayol, L., and Geiduschek, E.P. (1998) Affinity, stability and polarity of binding of the TATA binding protein governed by flexure at the TATA Box. *J Mol Biol* **282**: 731-739.

Grundy, W.N., Bailey, T.L., Elkan, C.P., and Baker, M.E. (1997) Meta-MEME: motif-based hidden Markov models of protein families. *Comput Appl Biosci* **13**: 397-406.

Hampsey, M. (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol Mol Biol Rev* **62**: 465-503.

Hannenhalli, S., and Levy, S. (2001) Promoter prediction in the human genome. *Bioinformatics* **17 Suppl 1**: S90-96.

Hardison, R.C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* **16**: 369-372.

Harley, C.B., and Reynolds, R.P. (1987) Analysis of E. coli promoter sequences. *Nucleic Acids Res* **15**: 2343-2361.

Harrison, S.C. (1991) A structural taxonomy of DNA-binding domains. *Nature* **353**: 715-719.

Hassoun, M.H. (1995) Fundamentals of Artificial Neural Networks. Cambridge, Massachusetts: MIT Press.

Hausner, W., Wettach, J., Hethke, C., and Thomm, M. (1996) Two transcription factors related with the eucaryal transcription factors TATA-binding protein and

transcription factor IIB direct promoter recognition by an archaeal RNA polymerase. *J Biol Chem* **271**: 30144-30148.

Hawley, D.K., and McClure, W.R. (1983) Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Res* **11**: 2237-2255.

Hayes, W.S., and Borodovsky, M. (1998) Deriving ribosomal binding site (RBS) statistical models from unannotated DNA sequences and the use of the RBS model for N-terminal prediction. *Pac Symp Biocomput*: 279-290.

Hecker, M., Schumann, W., and Volker, U. (1996) Heat-shock and general stress response in Bacillus subtilis. *Mol Microbiol* **19**: 417-428.

Hecker, M., and Volker, U. (1998) Non-specific, general and multiple stress resistance of growth-restricted Bacillus subtilis cells by the expression of the sigmaB regulon. *Mol Microbiol* **29**: 1129-1136.

Helmann, J.D., and Chamberlin, M.J. (1988) Structure and function of bacterial sigma factors. *Annu Rev Biochem* **57**: 839-872.

Helmann, J.D., Wu, M.F., Kobel, P.A., Gamo, F.J., Wilson, M., Morshedi, M.M., Navre, M., and Paddon, C. (2001) Global transcriptional response of Bacillus subtilis to heat shock. *J Bacteriol* **183**: 7318-7328.

Hengge-Aronis, R. (2002) Signal transduction and regulatory mechanisms involved in control of the sigma(S) (RpoS) subunit of RNA polymerase. *Microbiol Mol Biol Rev* **66**: 373-395, table of contents.

Hertz, G.Z., Hartzell, G.W., 3rd, and Stormo, G.D. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput Appl Biosci* **6**: 81-92.

Hertz, G.Z., and Stormo, G.D. (1995) Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps. In *Proceedings of the Third International Conference on Bioinformatics and Genome Research*. Lim, H.A., and Cantor, C.R. (ed.): World Scientific Publishing Co., pp. 201-216.

Hertz, G.Z., and Stormo, G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**: 563-577.

Hirano, F., Tanaka, H., Hirano, Y., Hiramoto, M., Handa, H., Makino, I., and Scheidereit, C. (1998) Functional interference of Sp1 and NF-kappaB through the same DNA binding site. *Mol Cell Biol* **18**: 1266-1274.

Hollams, E.M., Giles, K.M., Thomson, A.M., and Leedman, P.J. (2002) MRNA stability and the control of gene expression: implications for human disease. *Neurochem Res* **27**: 957-980.

Horn, P.J., and Peterson, C.L. (2002) Molecular biology. Chromatin higher order folding--wrapping up transcription. *Science* **297**: 1824-1827.

Horwitz, M.S., and Loeb, L.A. (1986) Promoters selected from random DNA sequences. *Proc Natl Acad Sci U S A* **83**: 7405-7409.

Hsu, L.M. (2002) Promoter clearance and escape in prokaryotes. *Biochim Biophys Acta* **1577**: 191-207.

Hu, Y.J., Sandmeyer, S., McLaughlin, C., and Kibler, D. (2000) Combinatorial motif analysis and hypothesis generation on a genomic scale. *Bioinformatics* **16**: 222-232.

Huang, X., and Helmann, J.D. (1998) Identification of target promoters for the Bacillus subtilis sigma X factor using a consensus-directed search. *J Mol Biol* **279**: 165-173.

Huang, X., Gaballa, A., Cao, M., and Helmann, J.D. (1999) Identification of target promoters for the Bacillus subtilis extracytoplasmic function sigma factor, sigma W. *Mol Microbiol* **31**: 361-371.

Huang, Y. (2002) Transcriptional silencing in Saccharomyces cerevisiae and Schizosaccharomyces pombe. *Nucleic Acids Res* **30**: 1465-1482.

Huet, J., Schnabel, R., Sentenac, A., and Zillig, W. (1983) Archaebacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type. *Embo J* **2**: 1291-1294.

Huffman, J.L., and Brennan, R.G. (2002) Prokaryotic transcription regulators: more than just the helix-turn-helix motif. *Curr Opin Struct Biol* **12**: 98-106.

Hunter, C.A. (1993) Sequence-dependent DNA structure. The role of base stacking interactions. *J Mol Biol* **230**: 1025-1054.

Hutchinson, G.B. (1996) The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput Appl Biosci* **12**: 391-398.

Ishihama, A., Kimura, M., and Mitsuzawa, H. (1998) Subunits of yeast RNA polymerases: structure and function. *Curr Opin Microbiol* **1**: 190-196.

Jeeves, M., Evans, P.D., Parslow, R.A., Jaseja, M., and Hyde, E.I. (1999) Studies of the Escherichia coli Trp repressor binding to its five operators and to variant operator

sequences. *Eur J Biochem* **265**: 919-928.

Jensen, L.J., and Knudsen, S. (2000) Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics* **16**: 326-333.

Jones, P.A., and Laird, P.W. (1999) Cancer epigenetics comes of age. *Nat Genet* **21**: 163-167.

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**: 195-202.

Jordan, M.I. (1999) Recurrent Networks. In *The MIT Encyclopedia of the Cognitive Sciences*. Wilson, R.A., Keil, F.C. (ed). Cambridge, Massachusetts: MIT Press.

Jordan, I.K., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. (2002) Microevolutionary genomics of bacteria. *Theor Popul Biol* **61**: 435-447.

Juo, Z.S., Chiu, T.K., Leiberman, P.M., Baikalov, I., Berk, A.J., and Dickerson, R.E. (1996) How proteins recognize the TATA box. *J Mol Biol* **261**: 239-254.

Kadam, S., and Emerson, B.M. (2002) Mechanisms of chromatin assembly and transcription. *Curr Opin Cell Biol* **14**: 262-268.

Kallipolitis, B.H., and Valentin-Hansen, P. (1998) Transcription of rpoH, encoding the Escherichia coli heat-shock regulator sigma32, is negatively controlled by the cAMP-CRP/CytR nucleoprotein complex. *Mol Microbiol* **29**: 1091-1099.

Karas, H., Knuppel, R., Schulz, W., Sklenar, H., and Wingender, E. (1996) Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. *Comput Appl Biosci* **12**: 441-446.

Kasabov, N.K. (1996) Foundations of neural networks, fuzzy systems, and knowledge. Boulder, Colorado: Bradford.

Katayama, A., Fujita, N., and Ishihama, A. (2000) Mapping of subunit-subunit contact surfaces on the beta' subunit of Escherichia coli RNA polymerase. *J Biol Chem* **275**: 3583-3592.

Keener, J., and Nomura, M. (1996) Regulation of Ribosome Synthesis. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. Vol. 1. Neidhardt, F.C. (ed). Washington, D.C.: American Society for Microbiology, pp. 1417-1431.

Keich, U., and Pevzner, P.A. (2002) Finding motifs in the twilight zone. *Bioinformatics* **18**: 1374-1381.

Kielbasa, S.M., Korbel, J.O., Beule, D., Schuchhardt, J., and Herzel, H. (2001) Combining frequency and positional information to predict transcription factor binding sites. *Bioinformatics* **17**: 1019-1026.

Kim, J.L., Nikolov, D.B., and Burley, S.K. (1993*a*) Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* **365**: 520-527.

Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. (1993*b*) Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365**: 512-520.

Kim, J.L., and Burley, S.K. (1994) 1.9 A resolution refined structure of TBP recognizing the minor groove of TATAAAAG. *Nat Struct Biol* **1**: 638-653.

Klingenhoff, A., Frech, K., Quandt, K., and Werner, T. (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* **15**: 180-186.

Knudsen, S. (1999) Promoter2.0: for the recognition of PolII promoter sequences. *Bioinformatics* **15**: 356-361.

Kohavi, R., and Provost, F. Glossary of terms. *Machine Learning* **30**: 271-274.

Kohonen, T. (1988) Self-organization and associative memory. New York: Springer-Verlag.

Kojima, C., Ulyanov, N.B., Kainosho, M., and James, T.L. (2001) Slow motion in the CAA*TTG sequence of a DNA decamer duplex studied by NMR. *Biochemistry* **40**: 7239-7246.

Konig, P., and Richmond, T.J. (1993) The X-ray structure of the GCN4-bZIP bound to ATF/CREB site DNA shows the complex depends on DNA flexibility. *J Mol Biol* **233**: 139-154.

Kosko, B. (1988) Bidirectional associative memories. *IEEE Transactions on Systems, Man, and Cybernetics* **18**: 49-60.

Koza, J. (1992) Genetic Programming. Cambridge, Massachusetts: MIT Press.

Kozak, M. (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene* **234**: 187-208.

Krakow, J.S., and Fronk, E. (1969) Azotobacter vinelandii ribonucleic acid polymerase. 8. Pyrophosphate exchange. *J Biol Chem* **244**: 5988-5993.

Krogh, A., Hertz, J.A. (1992) A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*. Lippman, D.S., Moody,

J.E., Touretzky, D.S. (ed): Morgan Kaufmann, pp. 950-957.

Kroos, L., and Yu, Y.T. (2000) Regulation of sigma factor activity during Bacillus subtilis development. *Curr Opin Microbiol* **3**: 553-560.

Kruger, K., Hermann, T., Armbruster, V., and Pfeifer, F. (1998) The transcriptional activator GvpE for the halobacterial gas vesicle genes resembles a basic region leucine-zipper regulatory protein. *J Mol Biol* **279**: 761-771.

Kuhnke, G., Theres, C., Fritz, H.J., and Ehring, R. (1989) RNA polymerase and gal repressor bind simultaneously and with DNA bending to the control region of the Escherichia coli galactose operon. *Embo J* **8**: 1247-1255.

Kuo, M.H., and Allis, C.D. (1998) Roles of histone acetyltransferases and deacetylases in gene regulation. *Bioessays* **20**: 615-626.

Kushner, S.R. (1996) mRNA Decay. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. Vol. 1. Neidhardt, F.C. (ed). Washington, D.C.: American Society for Microbiology, pp. 849-860.

Kyrpides, N.C., and Ouzounis, C.A. (1999) Transcription in archaea. *Proc Natl Acad Sci U S A* **96**: 8545-8550.

Lagrange, T., Kapanidis, A.N., Tang, H., Reinberg, D., and Ebright, R.H. (1998) New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev* **12**: 34-44.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L.,

Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., Szustakowki, J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., and Chen, Y.J. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Langer, D., Hain, J., Thuriaux, P., and Zillig, W. (1995) Transcription in archaea: similarity to that in eucarya. *Proc Natl Acad Sci U S A* **92**: 5768-5772.

Lawrence, C.E., and Reilly, A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **7**: 41-51.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**: 208-214.

Lawson, C.L., and Sigler, P.B. (1988) The structure of trp pseudorepressor at 1.65A shows why indole propionate acts as a trp 'inducer'. *Nature* **333**: 869-871.

Lee, T.I., and Young, R.A. (2000) Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* **34**: 77-137.

Leung, M.Y., Marsh, G.M., and Speed, T.P. (1996) Over- and underrepresentation of short DNA words in herpesvirus genomes. *J Comput Biol* **3**: 345-360.

Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G., and Lu, P. (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* **271**: 1247-1254.

Li, S., and Waters, R. (1998) Escherichia coli strains lacking protein HU are UV sensitive due to a role for HU in homologous recombination. *J Bacteriol* **180**: 3750-3756.

Li, H., Rhodius, V., Gross, C., and Siggia, E.D. (2002) Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc Natl Acad Sci U S A* **99**: 11772-11777.

Lisser, S., and Margalit, H. (1994) Determination of common structural features in Escherichia coli promoters by computer analysis. *Eur J Biochem* **223**: 823-830.

Liu, X., Brutlag, D.L., and Liu, J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*: 127-138.

Lohmann, R., Schneider, G., Behrens, D., and Wrede, P. (1994) A neural network model for the prediction of membrane-spanning amino acid sequences. *Protein Sci* **3**: 1597-1601.

Lopez-Garcia, P. (1999) DNA supercoiling and temperature adaptation: A clue to early diversification of life? *J Mol Evol* **49**: 439-452.

Luger, K., and Richmond, T.J. (1998*a*) DNA binding within the nucleosome core. *Curr Opin Struct Biol* **8**: 33-40.

Luger, K., and Richmond, T.J. (1998*b*) The histone tails of the nucleosome. *Curr Opin Genet Dev* **8**: 140-146.

Luscher, B. (2001) Function and regulation of the transcription factors of the Myc/Max/Mad network. *Gene* **277**: 1-14.

Maeda, H., Jishage, M., Nomura, T., Fujita, N., and Ishihama, A. (2000) Two extracytoplasmic function sigma subunits, sigma(E) and sigma(FecI), of Escherichia coli: promoter selectivity and intracellular levels. *J Bacteriol* **182**: 1181-1184.

Magill, C.P., Jackson, S.P., and Bell, S.D. (2001) Identification of a conserved archaeal RNA polymerase subunit contacted by the basal transcription factor TFB. *J Biol Chem* **276**: 46693-46696.

Mahadevan, I., and Ghosh, I. (1994) Analysis of E.coli promoter structures using neural networks. *Nucleic Acids Res* **22**: 2158-2165.

Mantovani, R. (1998) A survey of 178 NF-Y binding CCAAT boxes. *Nucleic Acids Res* **26**: 1135-1143.

Marchal, K., Thijs, G., De Keersmaecker, S., Monsieurs, P., De Moor, B., and Vanderleyden, J. (2003) Genome-specific higher-order background models to improve motif detection. *Trends Microbiol* **11**: 61-66.

Marilley, M., and Pasero, P. (1996) Common DNA structural features exhibited by eukaryotic ribosomal gene promoters. *Nucleic Acids Res* **24**: 2204-2211.

Marin, A., Gutierrez, G., and Oliver, J.L. (1999) Compositional correlation between open reading frames with opposite transcriptional orientations in Escherichia coli. *J Mol Evol* **48**: 712-716.

McAdams, H.H., and Arkin, A. (1997) Stochastic mechanisms in gene expression. *Proc Natl Acad Sci U S A* **94**: 814-819.

McDonnell, D.P., and Norris, J.D. (2002) Connections and regulation of the human estrogen receptor. *Science* **296**: 1642-1644.

McMahon, S.B., Wood, M.A., and Cole, M.D. (2000) The essential cofactor TRRAP recruits the histone acetyltransferase hGCN5 to c-Myc. *Mol Cell Biol* **20**: 556-562.

Medigue, C., Rose, M., Viari, A., and Danchin, A. (1999) Detecting and analyzing DNA sequencing errors: toward a higher quality of the Bacillus subtilis genome sequence. *Genome Res* **9**: 1116-1127.

Mehrotra, K., Mohan, C.K., Ranka, S. (1997) Elements of artificial neural networks. Boulder, Colorado: Bradford.

Mendez, R., Barnard, D., and Richter, J.D. (2002) Differential mRNA translation and meiotic progression require Cdc2-mediated CPEB destruction. *Embo J* **21**: 1833-1844.

Merika, M., and Thanos, D. (2001) Enhanceosomes. *Curr Opin Genet Dev* **11**: 205-208.

Minakhin, L., Bhagat, S., Brunning, A., Campbell, E.A., Darst, S.A., Ebright, R.H., and Severinov, K. (2001) Bacterial RNA polymerase subunit omega and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly. *Proc Natl Acad Sci U S A* **98**: 892-897.

Mironov, A.A., Koonin, E.V., Roytberg, M.A., and Gelfand, M.S. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res* **27**: 2981-2989.

Mitchell, M. (1998) Introduction to Genetic Algorithms. Cambridge, Massachusetts: MIT Press.

Monsalve, M., Mencia, M., Rojo, F., and Salas, M. (1996) Activation and repression of transcription at two different phage phi29 promoters are mediated by interaction of the same residues of regulatory protein p4 with RNA polymerase. *Embo J* **15**: 383-391.

Moran, C.P. (1993) RNA polymerase and transcription factors. In *Bacillus subtilis and Other Gram-Positive Bacteria: Biochemistry, Physiology, and Molecular Genetics*. Sonenshein, A.L., Hoch, J.A., Losick, R. (ed). Washington, D.C.: American Society for Microbiology, pp. 653-667.

Morita, M.T., Kanemori, M., Yanagi, H., and Yura, T. (2000) Dynamic interplay between antagonistic pathways controlling the sigma 32 level in Escherichia coli. *Proc Natl Acad Sci U S A* **97**: 5860-5865.

Müller-Hill, B. (1998) Some repressors of bacterial transcription. *Curr Opin Microbiol* **1**: 145-151.

Munn, M.M., and Alberts, B.M. (1991) DNA footprinting studies of the complex formed by the T4 DNA polymerase holoenzyme at a primer-template junction. *J Biol Chem* **266**: 20034-20044.

Murakami, K.S., Masuda, S., Campbell, E.A., Muzzin, O., and Darst, S.A. (2002) Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science* **296**: 1285-1290.

Murakami, K.S., Masuda, S., and Darst, S.A. (2002) Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 A resolution. *Science* **296**: 1280-1284.

Narayanan, A., Wu, X., and Yang, Z.R. (2002) Mining viral protease data to extract cleavage knowledge. *Bioinformatics* **18 Suppl 1**: S5-S13.

Naryshkin, N., Revyakin, A., Kim, Y., Mekler, V., and Ebright, R.H. (2000) Structural organization of the RNA polymerase-promoter open complex. *Cell* **101**: 601-611.

Natarajan, K., Jackson, B.M., Zhou, H., Winston, F., and Hinnebusch, A.G. (1999) Transcriptional activation by Gcn4p involves independent interactions with the SWI/SNF complex and the SRB/mediator. *Mol Cell* **4**: 657-664.

Needleman, S.B., and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443-453.

Neely, K.E., Hassan, A.H., Wallberg, A.E., Steger, D.J., Cairns, B.R., Wright, A.P., and Workman, J.L. (1999) Activation domain-mediated targeting of the SWI/SNF complex to promoters stimulates transcription from nucleosome arrays. *Mol Cell* **4**: 649-655.

Neidhardt, F.C., and Savageau, M.A. (1996) Regulation Beyond the Operon. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. Vol. 1. Neidhardt, F.C. (ed). Washington, D.C.: American Society for Microbiology, pp. 1310-1324.

Neidle, S. (1996) The crystallization and structure analysis of oligonucleotide sequences. In *Methods in Molecular Biology*. Vol. 56. Jones, C., Mulloy, B., Sanderson, M. (ed). Totowa, NJ: Humana Press Inc., pp. 267-292.

Neuhaus, D., Nakaseko, Y., Schwabe, J.W., and Klug, A. (1992) Solution structures of two zinc-finger domains from SWI5 obtained using two-dimensional 1H nuclear magnetic resonance spectroscopy. A zinc-finger structure with a third strand of beta-sheet. *J Mol Biol* **228**: 637-651.

Nomura, T., Fujita, N., and Ishihama, A. (1999) Mapping of subunit-subunit contact surfaces on the beta subunit of Escherichia coli RNA polymerase. *Biochemistry* **38**: 1346-1355.

Notredame, C., Higgins, D.G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205-217.

Nussinov, R. (1991) Signals in DNA sequences and their potential properties. *Comput Appl Biosci* **7**: 295-299.

Nussinov, R. (1992) The eukaryotic CCAAT and TATA boxes, DNA spacer flexibility and looping. *J Theor Biol* **155**: 243-270.

O'Donnell, S.M., and Janssen, G.R. (2001) The initiation codon affects ribosome binding and translational efficiency in Escherichia coli of cI mRNA with or without the 5' untranslated leader. *J Bacteriol* **183**: 1277-1283.

O'Neill, M.C. (1991) Training back-propagation neural networks to define and detect DNA-binding sites. *Nucleic Acids Res* **19**: 313-318.

O'Neill, M.C. (1992) Escherichia coli promoters: neural networks develop distinct descriptions in learning to search for promoters of different spacing classes. *Nucleic Acids Res* **20**: 3471-3477.

O'Shea, E.K., Klemm, J.D., Kim, P.S., and Alber, T. (1991) X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science* **254**: 539-544.

Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S., and Nishimura, Y. (1994) Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell* **79**: 639-648.

Ogata, K., Sato, K., and Tahirov, T. (2003) Eukaryotic transcriptional regulatory complexes: cooperativity from near and afar. *Curr Opin Struct Biol* **13**: 40-48.

Ohler, U., Harbeck, S., Niemann, H., Noth, E., and Reese, M.G. (1999) Interpolated markov chains for eukaryotic promoter recognition. *Bioinformatics* **15**: 362-369.

Ohler, U. (2000) Promoter prediction on a genomic scale--the Adh experience. *Genome Res* **10**: 539-542.

Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M., and Zhurkin, V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* **95**: 11163-11168.

Opel, M.L., Arfin, S.M., and Hatfield, G.W. (2001) The effects of DNA supercoiling on the expression of operons of the ilv regulon of Escherichia coli suggest a physiological rationale for divergently transcribed operons. *Mol Microbiol* **39**: 1109-1115.

Ozoline, O.N., Deev, A.A., and Arkhipova, M.V. (1997) Non-canonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by Escherichia coli RNA polymerase. *Nucleic Acids Res* **25**: 4703-4709.

Pabo, C.O., and Nekludova, L. (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J Mol Biol* **301**: 597-624.

Passner, J.M., and Steitz, T.A. (1997) The structure of a CAP-DNA complex having two cAMP molecules bound to each monomer. *Proc Natl Acad Sci U S A* **94**: 2843-2847.

Paszkowski, J., and Whitham, S.A. (2001) Gene silencing and DNA methylation processes. *Curr Opin Plant Biol* **4**: 123-129.

Paule, M.R., and White, R.J. (2000) Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res* **28**: 1283-1298.

Pedersen, A.G., and Engelbrecht, J. (1995) Investigations of Escherichia coli promoter sequences with artificial neural networks: new signals discovered upstream of the transcriptional startpoint. *Proc Int Conf Intell Syst Mol Biol* **3**: 292-299.

Pedersen, A.G., Baldi, P., Brunak, S., and Chauvin, Y. (1996) Characterization of

prokaryotic and eukaryotic promoters using hidden Markov models. *Proc Int Conf Intell Syst Mol Biol* **4**: 182-191.

Pedersen, A.G., Jensen, L.J., Brunak, S., Staerfeldt, H.H., and Ussery, D.W. (2000) A DNA structural atlas for Escherichia coli. *J Mol Biol* **299**: 907-930.

Perez-Martin, J., and Espinosa, M. (1994) Correlation between DNA bending and transcriptional activation at a plasmid promoter. *J Mol Biol* **241**: 7-17.

Perier, R.C., Praz, V., Junier, T., Bonnard, C., and Bucher, P. (2000) The eukaryotic promoter database (EPD). *Nucleic Acids Res* **28**: 302-303.

Pesole, G., Prunella, N., Liuni, S., Attimonelli, M., and Saccone, C. (1992) WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. *Nucleic Acids Res* **20**: 2871-2875.

Petersohn, A., Bernhardt, J., Gerth, U., Hoper, D., Koburger, T., Volker, U., and Hecker, M. (1999) Identification of sigma(B)-dependent genes in Bacillus subtilis using a promoter consensus-directed search and oligonucleotide hybridization. *J Bacteriol* **181**: 5718-5724.

Peterson, C.L., and Workman, J.L. (2000) Promoter targeting and chromatin remodeling by the SWI/SNF complex. *Curr Opin Genet Dev* **10**: 187-192.

Pettijohn, D.E. (1996) The Nucleoid. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. Vol. 1. Neidhardt, F.C. (ed). Washington, D.C.: American Society for Microbiology, pp. 158-166.

Pietrokovski, S., Henikoff, J.G., and Henikoff, S. (1996) The Blocks database--a system for protein classification. *Nucleic Acids Res* **24**: 197-200.

Ponomarenko, M.P., Ponomarenko, J.V., Kel, A.E., and Kolchanov, N.A. (1997) Search for DNA conformational features for functional sites. Investigation of the TATA box. *Pac Symp Biocomput*: 340-351.

Ponomarenko, M.P., Ponomarenko, J.V., Frolov, A.S., Podkolodny, N.L., Savinkova, L.K., Kolchanov, N.A., and Overton, G.C. (1999*a*) Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins. *Bioinformatics* **15**: 687-703.

Ponomarenko, M.P., Ponomarenko, J.V., Frolov, A.S., Podkolodnaya, O.A., Vorobyev, D.G., Kolchanov, N.A., and Overton, G.C. (1999*b*) Oligonucleotide frequency matrices addressed to recognizing functional DNA sites. *Bioinformatics* **15**: 631-643.

Prestridge, D.S., and Burks, C. (1993) The density of transcriptional elements in

promoter and non-promoter sequences. *Hum Mol Genet* **2**: 1449-1453.

Prestridge, D.S. (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol* **249**: 923-932.

Qian, Y.Q., Billeter, M., Otting, G., Muller, M., Gehring, W.J., and Wuthrich, K. (1989) The structure of the Antennapedia homeodomain determined by NMR spectroscopy in solution: comparison with prokaryotic repressors. *Cell* **59**: 573-580.

Qian, Y.Q., Resendez-Perez, D., Gehring, W.J., and Wuthrich, K. (1994) The des(1-6)antennapedia homeodomain: comparison of the NMR solution structure and the DNA-binding affinity with the intact Antennapedia homeodomain. *Proc Natl Acad Sci U S A* **91**: 4091-4095.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* **23**: 4878-4884.

Quandt, K., Grote, K., and Werner, T. (1996*a*) GenomeInspector: a new approach to detect correlation patterns of elements on genomic sequences. *Comput Appl Biosci* **12**: 405-413.

Quandt, K., Grote, K., and Werner, T. (1996*b*) GenomeInspector: basic software tools for analysis of spatial correlations between genomic structures within megabase sequences. *Genomics* **33**: 301-304.

Qureshi, S.A., and Jackson, S.P. (1998) Sequence-specific DNA binding by the S. shibatae TFIIB homolog, TFB, and its effect on promoter strength. *Mol Cell* **1**: 389-400.

Ramos, J.L., Gallegos, M.T., Marques, S., Ramos-Gonzalez, M.I., Espinosa-Urgel, M., and Segura, A. (2001) Responses of Gram-negative bacteria to certain environmental stressors. *Curr Opin Microbiol* **4**: 166-171.

Rashidzadeh, H., Khrapunov, S., Chance, M.R., and Brenowitz, M. (2003) Solution structure and interdomain interactions of the Saccharomyces cerevisiae "TATA binding protein" (TBP) probed by radiolytic protein footprinting. *Biochemistry* **42**: 3655-3665.

Ray, W.C., and Daniels, C.J. (2003) PACRAT: a database and analysis system for archaeal and bacterial intergenic sequence features. *Nucleic Acids Res* **31**: 109-113.

Record, M.T., Reznikoff, W.S., Craig, M.L., McQuade, K.L., Schlax, P.J. (1996) *Escherichia coli* RNA polymerase (Esigma70), promoters, and the kinetics of the

steps of transcription initiation. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. Neidhardt, F.C. (ed). Washington, D.C.: American Society for Microbiology, pp. 792-821.

Reed, R.D., Marks, R.J. (1999) Neural Smithing. Cambridge, Massachusetts: MIT Press.

Reese, M.G. (2001) Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. *Comput Chem* **26**: 51-56.

Reeve, J.N., Sandman, K., and Daniels, C.J. (1997) Archaeal histones, nucleosomes, and transcription initiation. *Cell* **89**: 999-1002.

Reitzer, L., and Schneider, B.L. (2001) Metabolic context and possible physiological themes of sigma(54)-dependent genes in Escherichia coli. *Microbiol Mol Biol Rev* **65**: 422-444, table of contents.

Reznikoff, W.S. (1992) The lactose operon-controlling elements: a complex paradigm. *Mol Microbiol* **6**: 2419-2422.

Rhodius, V.A., and Busby, S.J. (1998) Positive activation of gene expression. *Curr Opin Microbiol* **1**: 152-159.

Rice, P.A., Yang, S., Mizuuchi, K., and Nash, H.A. (1996) Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell* **87**: 1295-1306.

Richterich, P. (1998) Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res* **8**: 251-259.

Robertson, K.D. (2001) DNA methylation, methyltransferases, and cancer. *Oncogene* **20**: 3139-3155.

Robertson, K.D. (2002) DNA methylation and chromatin - unraveling the tangled web. *Oncogene* **21**: 5361-5379.

Robinson, K.A., and Lopes, J.M. (2000) SURVEY AND SUMMARY: Saccharomyces cerevisiae basic helix-loop-helix proteins regulate diverse biological processes. *Nucleic Acids Res* **28**: 1499-1505.

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.

Rogers, J. (1996) Object Oriented Neural Networks in C++. New York: Academic Press.

Rosen, R., and Ron, E.Z. (2002) Proteome analysis in the study of the bacterial heat-shock response. *Mass Spectrom Rev* **21**: 244-265.

Rosenberg, J.M., Seeman, N.C., Kim, J.J., Suddath, F.L., Nicholas, H.B., and Rich, A. (1973) Double helix at atomic resolution. *Nature* **243**: 150-154.

Rosenblatt, F. (1962) Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington, D.C.: Spartan.

Roy, B., and Lee, A.S. (1995) Transduction of calcium stress through interaction of the human transcription factor CBF with the proximal CCAAT regulatory element of the grp78/BiP promoter. *Mol Cell Biol* **15**: 2263-2274.

Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986) Learning representations by back-propagation errors. *Nature* **323**: 533-536.

Saier, M.H., Ramseier, T.M., Reizer, J. (1996) Regulation of carbon utilization. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. Vol. 1. Neidhardt, F.C. (ed). Washington, D.C.: American Society for Microbiology, pp. 1325-1343.

Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., and Collado-Vides, J. (2000) Operons in Escherichia coli: genomic analyses and predictions. *Proc Natl Acad Sci U S A* **97**: 6652-6657.

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate, D., Diaz-Peredo, E., Sanchez-Solano, F., Perez-Rueda, E., Bonavides-Martinez, C., and Collado-Vides, J. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Res* **29**: 72-74.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977) Nucliotide sequence of bacteriophage phi X174 DNA. *Nature* **265**: 687-695.

Satchwell, S.C., and Travers, A.A. (1989) Asymmetry and polarity of nucleosomes in chicken erythrocyte chromatin. *Embo J* **8**: 229-238.

Sato, S., Hagihara, M., Sugimoto, K., and Morii, T. (2002) Chemical approaches untangling sequence-specific DNA binding by proteins. *Chemistry* **8**: 5066-5071.

Schalkoff, R.J. (1992) Pattern Recognition: Statistical, Syntactic and Neural Approaches. New York: John Wiley and Sons.

Schbath, S. (1997) An efficient statistic to detect over- and under-represented words in DNA sequences. *J Comput Biol* **4**: 189-192.

Scherf, M., Klingenhoff, A., and Werner, T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel

context analysis approach. *J Mol Biol* **297**: 599-606.

Schleif, R. (2003) AraC protein: a love-hate relationship. *Bioessays* **25**: 274-282.

Schneider, G., and Wrede, P. (1994) The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. *Biophys J* **66**: 335-344.

Schneider, R., Travers, A., Kutateladze, T., and Muskhelishvili, G. (1999) A DNA architectural protein couples cellular physiology and DNA topology in Escherichia coli. *Mol Microbiol* **34**: 953-964.

Schneider, R., Travers, A., and Muskhelishvili, G. (2000) The expression of the Escherichia coli fis gene is strongly dependent on the superhelical density of DNA. *Mol Microbiol* **38**: 167-175.

Schnitzler, G., Sif, S., and Kingston, R.E. (1998) Human SWI/SNF interconverts a nucleosome between its base state and a stable remodeled state. *Cell* **94**: 17-27.

Schulz, A., Langowski, J., and Rippe, K. (2000) The effect of the DNA conformation on the rate of NtrC activated transcription of Escherichia coli RNA polymerase.sigma(54) holoenzyme. *J Mol Biol* **300**: 709-725.

Schumacher, M.A., Choi, K.Y., Zalkin, H., and Brennan, R.G. (1994) Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. *Science* **266**: 763-770.

Sekinger, E.A., and Gross, D.S. (2001) Silenced chromatin is permissive to activator binding and PIC recruitment. *Cell* **105**: 403-414.

Sentenac, A., Riva, M., Thuriaux, P., Buhler,J.-M., Treich, I., Carles, C., Werner, M., Ruet, A., Huet, J., Mann, C., Chiannikulchai, N., Stettler, S., Mariotte, S. (1992) Yeast RNA polymerase subunits and genes. In *Transcriptional Regulation*. McKnight, S.L., Yamamoto, K.R. (ed): Cold Spring Harbor Laboratory Press, pp. 27-54.

Shine, J., and Dalgarno, L. (1974) The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A* **71**: 1342-1346.

Sinden, R.R., and Pettijohn, D.E. (1981) Chromosomes in living Escherichia coli cells are segregated into domains of supercoiling. *Proc Natl Acad Sci U S A* **78**: 224-228.

Skovgaard, M., Jensen, L.J., Brunak, S., Ussery, D., and Krogh, A. (2001) On the total number of genes and their length distribution in complete microbial genomes.

Trends Genet **17**: 425-428.

Slupska, M.M., King, A.G., Fitz-Gibbon, S., Besemer, J., Borodovsky, M., and Miller, J.H. (2001) Leaderless transcripts of the crenarchaeal hyperthermophile Pyrobaculum aerophilum. *J Mol Biol* **309**: 347-360.

Smith, T.F., and Waterman, M.S. (1981) Identification of common molecular subsequences. *J Mol Biol* **147**: 195-197.

Sonnhammer, E.L., von Heijne, G., and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6**: 175-182.

Soppa, J. (1999) Transcription initiation in Archaea: facts, factors and future aspects. *Mol Microbiol* **31**: 1295-1305.

Steffen, N.R., Murphy, S.D., Tolleri, L., Hatfield, G.W., and Lathrop, R.H. (2002) DNA sequence and structure: direct and indirect recognition in protein-DNA binding. *Bioinformatics* **18 Suppl 1**: S22-S30.

Stormo, G.D., Schneider, T.D., and Gold, L.M. (1982) Characterization of translational initiation sites in E. coli. *Nucleic Acids Res* **10**: 2971-2996.

Stormo, G.D., and Hartzell, G.W., 3rd (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci U S A* **86**: 1183-1187.

Stormo, G.D. (1990) Consensus patterns in DNA. *Methods Enzymol* **183**: 211-221.

Stormo, G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics* **16**: 16-23.

Strahl, B.D., and Allis, C.D. (2000) The language of covalent histone modifications. *Nature* **403**: 41-45.

Straney, S.B., and Crothers, D.M. (1987) Lac repressor is a transient gene-activating protein. *Cell* **51**: 699-707.

Struhl, K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98**: 1-4.

Stuger, R., Woldringh, C.L., van der Weijden, C.C., Vischer, N.O., Bakker, B.M., van Spanning, R.J., Snoep, J.L., and Westerhoff, H.V. (2002) DNA supercoiling by gyrase is linked to nucleoid compaction. *Mol Biol Rep* **29**: 79-82.

Sudarsanam, P., and Winston, F. (2000) The Swi/Snf family nucleosome-remodeling complexes and transcriptional control. *Trends Genet* **16**: 345-351.

Sugimoto, N., Nakano, S., Yoneyama, M., and Honda, K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res* **24**: 4501-4505.

Susa, M., Sen, R., and Shimamoto, N. (2002) Generality of the branched pathway in transcription initiation by Escherichia coli RNA polymerase. *J Biol Chem* **277**: 15407-15412.

Sutton, R.S., Barto, A.G. (1998) Reinforcement Learning: An Introduction. Cambridge, Massachusetts: MIT Press.

Tam, C., Collinet, B., Lau, G., Raina, S., and Missiakas, D. (2002) Interaction of the conserved region 4.2 of sigma(E) with the RseA anti-sigma factor. *J Biol Chem* **277**: 27282-27287.

Teng, Y., Yu, S., and Waters, R. (2001) The mapping of nucleosomes and regulatory protein binding sites at the Saccharomyces cerevisiae MFA2 gene: a high resolution approach. *Nucleic Acids Res* **29**: E64-64.

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.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17**: 1113-1122.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673-4680.

Thompson, D.K., and Daniels, C.J. (1998) Heat shock inducibility of an archaeal TATA-like promoter is controlled by adjacent sequence elements. *Mol Microbiol* **27**: 541-551.

Thompson, J., van Spaendonk, R.M., Choudhuri, R., Sinden, R.E., Janse, C.J., and Waters, A.P. (1999) Heterogeneous ribosome populations are present in Plasmodium berghei during development in its vector. *Mol Microbiol* **31**: 253-260.

Travers, A. (1997) DNA-protein interactions: IHF--the master bender. *Curr Biol* **7**: R252-254.

van Helden, J., Andre, B., and Collado-Vides, J. (1998) Extracting regulatory sites from

the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* **281**: 827-842.

van Helden, J., Rios, A.F., and Collado-Vides, J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* **28**: 1808-1818.

Vanet, A., Marsan, L., Labigne, A., and Sagot, M.F. (2000) Inferring regulatory elements from a whole genome. An analysis of Helicobacter pylori sigma(80) family of promoter signals. *J Mol Biol* **297**: 335-353.

Vo, N.V., Hsu, L.M., Kane, C.M., and Chamberlin, M.J. (2003) In vitro studies of transcript initiation by Escherichia coli RNA polymerase. 3. Influences of individual DNA elements within the promoter recognition region on abortive initiation and promoter escape. *Biochemistry* **42**: 3798-3811.

Vologodskaia, M., and Vologodskii, A. (2002) Contribution of the intrinsic curvature to measured DNA persistence length. *J Mol Biol* **317**: 205-213.

Vuister, G.W., Kim, S.J., Wu, C., and Bax, A. (1994) NMR evidence for similarities between the DNA-binding regions of Drosophila melanogaster heat shock factor and the helix-turn-helix and HNF-3/forkhead families of transcription factors. *Biochemistry* **33**: 10-16.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.L. (1989) Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustic, Speech, and Signal Processing* **37**: 328-339.

Wang, P., Yang, J., Lawley, B., and Pittard, A.J. (1997) Repression of the aroP gene of Escherichia coli involves activation of a divergent promoter. *J Bacteriol* **179**: 4213-4218.

Wang, H.C., Badger, J., Kearney, P., and Li, M. (2001) Analysis of codon usage patterns of bacterial genomes using the self-organizing map. *Mol Biol Evol* **18**: 792-800.

Weickert, M.J., and Adhya, S. (1993) The galactose regulon of Escherichia coli. *Mol Microbiol* **10**: 245-251.

Werel, W., Schickor, P., and Heumann, H. (1991) Flexibility of the DNA enhances promoter affinity of Escherichia coli RNA polymerase. *Embo J* **10**: 2589-2594.

Werner, F., Eloranta, J.J., and Weinzierl, R.O. (2000) Archaeal RNA polymerase subunits F and P are bona fide homologs of eukaryotic RPB4 and RPB12. *Nucleic Acids Res* **28**: 4299-4305.

White, M.F., and Bell, S.D. (2002) Holding it together: chromatin in the Archaea. *Trends*

Genet **18**: 621-626.

Wing, R., Drew, H., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R.E. (1980) Crystal structure analysis of a complete turn of B-DNA. *Nature* **287**: 755-758.

Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**: 238-241.

Wintjens, R., and Rooman, M. (1996) Structural classification of HTH DNA-binding domains and protein-DNA interaction modes. *J Mol Biol* **262**: 294-313.

Wojtuszewski, K., Hawkins, M.E., Cole, J.L., and Mukerji, I. (2001) HU binding to DNA: evidence for multiple complex formation and DNA bending. *Biochemistry* **40**: 2588-2598.

Wolfe, S.A., Nekludova, L., and Pabo, C.O. (2000) DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct* **29**: 183-212.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin, A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris, D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E.J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., McDonald, S., McLean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell, C., Oliver, K., O'Neil, S., Pearson, D., Quail, M.A., Rabbinowitsch, E., Rutherford, K., Rutter, S., Saunders, D., Seeger, K., Sharp, S., Skelton, J., Simmonds, M., Squares, R., Squares, S., Stevens, K., Taylor, K., Taylor, R.G., Tivey, A., Walsh, S., Warren, T., Whitehead, S., Woodward, J., Volckaert, G., Aert, R., Robben, J., Grymonprez, B., Weltjens, I., Vanstreels, E., Rieger, M., Schafer, M., Muller-Auer, S., Gabel, C., Fuchs, M., Dusterhoft, A., Fritzc, C., Holzer, E., Moestl, D., Hilbert, H., Borzym, K., Langer, I., Beck, A., Lehrach, H., Reinhardt, R., Pohl, T.M., Eger, P., Zimmermann, W., Wedler, H., Wambutt, R., Purnelle, B., Goffeau, A., Cadieu, E., Dreano, S., Gloux, S., Lelaure, V., Mottier, S., Galibert, F., Aves, S.J., Xiang, Z., Hunt, C., Moore, K., Hurst, S.M., Lucas, M., Rochet, M., Gaillardin, C., Tallada, V.A., Garzon, A., Thode, G., Daga, R.R., Cruzado, L., Jimenez, J., Sanchez, M., del Rey, F., Benito, J., Dominguez, A., Revuelta, J.L., Moreno, S., Armstrong, J., Forsburg, S.L., Cerutti, L., Lowe, T., McCombie, W.R., Paulsen, I., Potashkin, J., Shpakovski, G.V., Ussery, D., Barrell, B.G., Nurse, P., and Cerrutti, L. (2002) The genome sequence of Schizosaccharomyces pombe. *Nature* **415**: 871-880.

Workman, C.T., and Stormo, G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput*: 467-478.

Wosten, M.M. (1998) Eubacterial sigma-factors. *FEMS Microbiol Rev* **22**: 127-150.

Wu, T.H., Grelland, E., Boye, E., and Marinus, M.G. (1992) Identification of a weak promoter for the dam gene of Escherichia coli. *Biochim Biophys Acta* **1131**: 47-52.

Wu, C.H. (1997) Artificial neural networks for molecular sequence analysis. *Comput Chem* **21**: 237-256.

Xu, H., and Hoover, T.R. (2001) Transcriptional regulation at a distance in bacteria. *Curr Opin Microbiol* **4**: 138-144.

Yada, T., Sazuka, T., and Hirosawa, M. (1997) Analysis of sequence patterns surrounding the translation initiation sites on Cyanobacterium genome using the hidden Markov model. *DNA Res* **4**: 1-7.

Yanagi, K., Prive, G.G., and Dickerson, R.E. (1991) Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *J Mol Biol* **217**: 201-214.

Yang, J., Wang, P., and Pittard, A.J. (1999) Mechanism of repression of the aroP P2 promoter by the TyrR protein of Escherichia coli. *J Bacteriol* **181**: 6411-6418.

Yanofsky, C. (2000) Transcription attenuation: once viewed as a novel regulatory strategy. *J Bacteriol* **182**: 1-8.

Yoder, J.A., Walsh, C.P., and Bestor, T.H. (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* **13**: 335-340.

Young, B.A., Gruber, T.M., and Gross, C.A. (2002) Views of transcription initiation. *Cell* **109**: 417-420.

Yudkovsky, N., Logie, C., Hahn, S., and Peterson, C.L. (1999) Recruitment of the SWI/SNF chromatin remodeling complex by transcriptional activators. *Genes Dev* **13**: 2369-2374.

Zak, M. (1989) Terminal attractors in neural networks. *Neural Networks* **2**: 259-274.

Zar, J.H. (1999) Biostatstical Analysis. Upper Saddle River, NJ: Prentice Hall.

Zemzoumi, K., Frontini, M., Bellorini, M., and Mantovani, R. (1999) NF-Y histone fold alpha1 helices help impart CCAAT specificity. *J Mol Biol* **286**: 327-337.

Zhang, M.Q. (1998) Identification of human gene core promoters in silico. *Genome Res* **8**: 319-326.

Zhang, M.Q. (2000) Discriminant analysis and its application in DNA sequence motif recognition. *Brief Bioinform* **1**: 331-342.