IDENTIFICATION OF CIS-ACTING REGULATORY ELEMENTS USING ORTHOLOGY BIASED GIBBS SAMPLING

by

Keith Anthony Boroevich
BSc, Simon Fraser University, 2002

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the
Department
of
Molecular Biology and Biochemistry

SIMON FRASER UNIVERSITY

Fall 2005

APPROVAL

Name: **Keith Anthony Boroevich**

Degree: **Master of Science**

Title of Thesis: **Identification of cis-acting regulatory elements using orthology biased gibbs sampling**

Examining Committee:

Chair: **Dr. E. Emberly**
Assistant Professor
Department of Physics

____

Dr. D. L. Baillie
Co-Senior Supervisor
Professor
Department of Molecular Biology and Biochemistry

____

Dr. S.J.M. Jones
Co-Senior Supervisor
Assistant Professor
Department of Medical Genetics, University of British Columbia

____

Dr. F. Pio
Supervisory Committee member
Assistant Professor
Department of Molecular Biology and Biochemistry

____

Dr. A. Beckenbach
Internal Examiner
Professor
Department of Biology

Date Defended/Approved: October 14, 2005

DECLARATION OF PARTIAL COPYRIGHT LICENCE

ABSTRACT

Many computational pattern searching tools for the discovery of novel, common regulatory elements between co-expressed genes have been developed over the last ten years. However, few approaches attempt to incorporate valuable additional information, such as inter-species conservation, into the prediction process. Orthology biased Gibbs sampler (OrBS) is an expansion on the Gibbs sampler motif discovery approach. I introduce dynamic motif width prediction, a novel convergence detection approach, and a scoring function that incorporates cross-species sequence conservation. The algorithm was refined using the *Caenorhabditis elegans* X box element and is shown to successfully identify the element in sequence sets with only 33% of X box regulated genes. Using the reported X box consensus, I successfully identified additional genes, like the *C. elegans* orthologue to human BBS4. OrBS was less successful in the identification of the other *C. elegans* regulatory elements, such as the PHA-4 binding site and the UNC-86 binding site.

KEYWORDS

cis-regulatory element, regulatory element detection, Gibbs sampling, computational biology, transcriptional control, *Caenorhabditis elegans*,

ACKNOWLEDGEMENTS

I would like to acknowledge my Senior Supervisors, Dr. David Baillie and Dr. Steven Jones, for their insight and guidance throughout graduate school. I acknowledge the remainder of my supervisory committee, Dr. Frederic Pio, for his advice and support. I would also like to thanks Dr. Dixie Mager, Dr. Michel Leroux and Dr. Oliver Blacque for giving me the opportunity to broaden my research experience.

Thanks are due to all members of the Baillie Lab, specifically Dr. Robert Johnsen, Allan Mah, Domena Tu, Carrie Simms, Matrin Jones, and Heesun Shin, for their advice, encouragement, and levity.

I am indebted to Duncan Napier, the systems and instrumentation consultant for the Department of Molecular Biology and Biochemistry, and Mathew Laid, of Dr. Fiona Brinkman's Laboratory, for assistance with computational problems and advice with programming strategies.

A large portion of the funding of this project is from the Canadian Institutes of Health Research (CIHR) and the Michael Smith Foundation for Health Research (MSFHR).

Last but now least, I acknowledge the other students of the CIHR/MSFHR Strategic Training Program in Bioinformatics and wish them success in their future endeavours.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The quest for determining what constitutes and causes a given gene's particular expression pattern has long been a matter of research. Many expression differences between genes can be explained by the existence of *cis*-acting regulatory elements that modulate the transcription of a gene, through binding of a protein. In addition to a host of other mechanisms that can regulate expression, a transcription factor binding site (TFBS) can be located either upstream or downstream of a gene or within an intron of the gene, and can exist a great distance from the gene. In combination with the small size of the elements, typically 6 to 12 nucleotides, the discovery and identification of *cis*-acting regulatory elements is a difficult task. Traditional genetic and molecular biology techniques for the discovery of a novel TFBS are time and labour intensive. Once the affected gene of a transcription factor is identified through genetic experimentation, a handful of procedures are available to locate the regulatory element: footprinting uses a chemical or nuclease to degrade a DNA strand in areas the protein does not bind; linker scanning mutagenesis can localize elements through replacing segments of the promoter with a synthetic linker of equal length using restriction enzymes (McKnight and Kingsbury 1982); gel mobility shift assays can identify protein-DNA interacting through mobility retardation on an electrophoresis gel. While through these techniques, the TFBS discovery problem has been partially solved for many genes individually, undiscovered TFBSs, methodologies to identify novel TFBSs in a high throughput manner have only come about within the last decade.

Most of these methods are computationally based and owe their emergence to the ongoing rise in computational power and decrease in sequencing costs. One of the common computational prediction methods, and the method applied in my Orthology Biased Gibbs Sampler (OrBS), is the Gibbs sampling approach, a Markov Chain Monte Carlo (MCMC), method which gained much popularity in the 1980s in the field of image processing (German and German 1984). A Markov chain consists of a finite number of states. At each 'time' interval, a Markov chain process must occupy one, and only one, of these states. As time advances, the process either stays in its current state, or moves to another state. The movement of a Markov chain process is dependent on a defined set of transmission probabilities. In addition, a Markov chain must abide by the following two characteristics: the *memoryless property* which states that the probability of the next state depends only on the current state and no past states, and the *time homogeneity property* which states the probability of the next state is independent of time (Ewens and Grant 2001). The second half of the description, Monte Carlo, refers to the stochastic nature of the method. Monte Carlo methods are typically used to simulate an unknown target distribution by sampling randomly from this distribution. After many iterative steps, if the sample distribution reaches equilibrium, i.e. changes insignificantly after each new sample, it is assumed to simulate the target distribution. MCMC methods sample from an unknown target distribution by randomly 'stepping' through states at each time interval. The larger the number of steps taken, the better the sample distribution approximates the target distribution.

The first successful attempt of the application of Gibbs sampling to the identification of novel biological motifs is the site sampler, or Gibbs sampler (Lawrence *et al.* 1993). In the field of alphabetic motif recognition, including protein and nucleotide motifs, the site sampler is a heuristic multiple local alignment program. Starting with a pattern created from randomly selected motifs from the user provided sequence set, the site sampler iteratively refines the pattern through the maximization of a scoring function. Typically the scoring function is based on the probability of a motif belonging to the pattern over the probability that same motif belongs to the sequence background. New motifs are sampled from each sequence based on the probability distribution created from the scoring function. The site sampler does have its drawbacks. All sequences in the sequence set are assumed to contain a user specified number of instances of the unknown motif, thus its predictive power was greatly dependant on the ability of the user to construct a biologically correct input sequence set. A later incarnation by the same research group, the motif sampler, addressed this problem by sampling each sequence position individually (Neuwald *et al.* 1995). Each position was then assigned to the motif model or the background model. Although an improvement, this biases the algorithm to discover more motifs in sequences of greater length. Another issue is the width of the motif pattern is static, requiring the user to guess the unknown motif width. Other drawbacks of this program exist and are addressed later in this paper.

Despite the preceding, the site sampler not only laid the ground work for later, more elaborate algorithms, but validates the application of general pattern recognition algorithms to biology. It was shown to successfully identify protein motifs in helix-turn-helix proteins, lipocalins, and prenyltransferases, given prior knowledge of motif width and motif occurrence(Lawrence *et al.* 1993). This shows that with little modification of the basic algorithm, Gibbs sampling is quite capable of modelling and identifying biological motifs.

Gibbs sampling is not the only statistical approach to the identification of TFBSs.

Over the past ten years many algorithms have been developed and refined. Some of the most popular and innovative programs are: MotifSampler, a recent adaptation of the Gibbs sampling approach (Thijs *et al.* 2002); Consensus and MEME, energy maximization (EM) algorithms which maximize the motifs information content (IC) or a related measure (Bailey and Elkan 1995b; Hertz and Stormo 1999); and Weeder, an exhaustive word-counting algorithm with substitution allowance (Pavesi *et al.* 2004).

The MotifSampler is a recent incarnation of the Gibbs sampling algorithm, the site sampler, previously discussed. It expands this algorithm in two major ways. The first major expansion is the use of a probability distribution to estimate the expected number of occurrences of a motif in a sequence. The occurrence estimation algorithm first estimates the probability of observing zero motifs in a sequence, or in other words, the probability the sequence was drawn from the background model. It then iteratively estimates the probability of observing an increasing number of motifs in the up to a maximum occurrence. In order to efficiently calculate these probabilities, MotifSampler employs a modified *forward* algorithm, commonly used to compute the likelihood a given sequence is drawn from a Hidden Markov Model (HMM). From this occurrence probability distribution, an expected occurrence can be calculated. A more detailed explanation of this algorithm is provided in the Methods and Materials section.

The second major expansion of the basic Gibbs sampling algorithm is the use of a higher order background model. The site sampler creates the background model from the input sequence set. The authors adopt this methodology from other computational prediction tools, such as gene finding programs, and go on to show that this alone can increase the predictive ability of the Gibbs sampling technique (Thijs *et al.* 2001). The results show that 3rd-order or 4th-order models are most effective.

The MotifSampler exhibits varying success in the identification of known TFBSs. The program was highly successful in identifying the G-box, a ubiquitous regulatory element found in plant genomes that is bound by the GBF (G-box binding factors) family members (Menkens *et al.* 1995), and the FNR (fumarate, nitrate reduction) binding site of bacterial genes. However, only moderate success was observed in the identification of the promoter elements of methionine response gene in *S. cerevisiae*, where known motifs were observed in approximately 50% of the runs.

The EM algorithms, such as Consensus and MEME, are *greedy* algorithms. Greedy algorithms attempt to find the global optimum to a problem by making the locally optimum choice at each stage. The basic algorithm of Consensus is as follows. First, all valid motif positions of a selected sequence are assessed and a motif model is created for each. At this point, all motif models are considered 'interesting'. The next sequence is then added to the analysis and all valid motif positions are compared to the interesting motif models. Only those comparisons that pass some criterion, for example a minimum IC score or the top *X* number of comparisons, are used to create a new set of interesting motif models. This is repeated until all sequences are added to the analysis resulting in a final set of interesting motif models that are reported back to the user. MEME uses a slightly different algorithm, where a single EM run is executed for all possible motif positions in the entire sequence set (Bailey and Elkan 1995a). The highest scoring motif is then chosen and run to completion, resulting in a single prediction. The most interesting aspects of MEME are the refinements and expansions to the program. The current version of MEME uses amino acid chemical property similarities in calculating a motif score and is able to specifically identify palindromic DNA motifs (Bailey and Elkan 1995b).

One major drawback of the EM algorithm of Consensus arises if the initial sequences added to the analysis lack or have a weak consensus to some true motif, that element will not be discovered. In order to avoid this, the authors suggest that the program is run multiple times, thus randomizing the sequence analysis order and minimizing the chances of missing an optimum.

Recently, MEME was used to identify a putative Arc-A binding site in *Escherichia coli* K12 (Salmon *et al.* 2005). Arc-A is involved in the regulation of aerobic/anaerobic expression of genes. The gene set was created from the results of a DNA microarray analysis of the transition of *E. coli* from an aerobic state to an anaerobic one (Salmon *et al.* 2003). Unfortunately, like most results derived from motif discovery tools, while the putative motif was used to identify an additional set of coregulated genes, it was not experimentally validated as a regulatory element.

The methodology applied in Weeder is the most unique and exhaustive of the tools discussed so far. In essence, the program examines all oligos of a given length in the sequence set and reports those oligos that are overrepresented. While being, perhaps, the most intuitive methodology, this *exact* approach is also the most difficult to implement efficiently. Examination of all possible oligos is extremely time consuming and only feasible for motifs of a small length. Additionally, a strict comparison approach does not allow for the natural degeneracy of most regulatory elements. To overcome this problem, Weeder starts with an exact approach, using suffix trees to generate an oligo database. Then Weeder attempt to identify groups oligos of a length, *M*, that differ within the group by a maximum of *e* mutations, and contain at least *q* members. This is accomplished by first examining groups of oligos of a shorter length, *m*, that conform to the same mutation and membership rules. The authors state that the algorithm can miss significant motifs based in the limited search space and show no results of real-data applications. However, in a recent comparison of 13 regulatory element discovery tools, Weeder outperformed all other tools in most measures of predictive power (Tompa *et al.* 2005). Despite this, at the time of writing this, there has been no application of this algorithm to real problems.

Out of the three main methodologies for computational motif discovery, the Gibbs sampling approach was selected. The first reason is based on the expansion of the algorithms to incorporate prior knowledge. The current approaches of computational motif discovery fall into two categories: the analysis of co-expressed genes (intraspecies) and the analysis of orthologous genes (interspecies). It is our goal to merge these two approaches into one biologically meaningful approach. Some researchers have simply merged such data into a single sequence set, however there are problems with such an approach. While coregulated genes in a single species are all regulated by the same transcription factors (TFs) by definition, orthologous genes are, at best, regulated by orthologous TFs, with possibly different DNA binding affinities. If this were the case, although the regulatory motifs that each respective TF binds might appear similar, the resulting diagnostic motif built from these motifs would be incorrect for both species. Instead, we use the information from interspecies comparisons to bias the program towards motifs with high interspecies conservation while maintaining the species specificity of the resulting motif model. Inclusion of such non-sequence information is not possible in the word-counting methodology of Weeder.

The decision between Gibbs sampling and EM algorithms is a more subtle. Both approaches are capable of being expanded to include prior information. However, the application of a greedy algorithm to a problem that cannot be directly solved by it is offsetting. The Monte Carlo aspect of the Gibb sampling approach allows the program to jump between local maxima in search of the global maximum and not get stalled in them. In order to test and refine OrBS, a set of test data is required. The roundworm, *Caenorhabditis elegans* was chosen as the model organism for this project. Nematoda has always been a phylum of much research due to the many species parasitic to humans. Introduction of nematodes, namely *C. elegans* and *Caenorhabditis briggsae*, into genetic research can be accredited to Ellsworth C. Dougherty (Dougherty and Calhoun 1949), who was interested in the nutritional requirements and axenic cultivation of the nematode. Later he provided Nobel laureate Sydney Brenner with a culture of *C. elegans*. Brenner was interested in the genetics of behaviour and believed that *C. elegans* would make a better model organism than Drosophila because of the much smaller size of nervous system (Brenner 1974).

There are multiple reasons that make *C. elegans* amenable to research in regulatory element discovery. Perhaps the most beneficial is the relative ease with which transgenic individuals can be made (Fire 1986). Also, the fate of all cell lineages is *C. elegans* has been mapped (Sulston 2003; Sulston and Horvitz 1977; Sulston *et al.* 1983). An adult hermaphrodite worm has 959 somatic cells, all of which have been identified by the location of their nuclei. The natural translucency of the organism allows *in vivo* analysis of gene expression via transgenic strains in which the promoter region of the gene of interest is fused with suitable marker gene. Such expression mapping has been the work of the *Caenorhabditis elegans* Gene Expression Project (*Ce*GEP), of which the Baillie Laboratory at Simon Fraser University is a part (McKay *et al.* 2003). Site directed mutagenesis of such constructs can be used to experimentally validate any predicted regulatory elements.

Some of the first regulatory elements to be resolved in *C. elegans* are the vitellogenins regulatory elements, VPE1 and VPE2. Vitellogenins are developmentally regulated genes typically only expressed in the female of the species and are regulated by the hormone estrogen (MacMorris *et al.* 1992). In *C. elegans*, the six vitellogenin gene family members are expressed only in late L4 and adult hermaphrodite worms and only in the intestine (Blumenthal *et al.* 1984). The sex, stage, and tissue specific expression pattern and strong similarity to the vertebrate vitellogenin family made these genes interesting candidates for study of transcriptional regulatory mechanisms (MacMorris *et al.* 1992). The first step into the discovery of the mechanisms of vitellogenin regulatory controls was the creation of a transgenic worm carrying a *vit-2*::*vit-6* fusion protein. The construct, which was detected immunologically and by nuclease protection, confirmed the expression pattern of the vitellogenins and resolved the possible location of *cis*-acting regulatory elements to 3.9Kb upstream of *vit-2* and 600bp downstream of *vit-6* (Riddle 1997; Spieth *et al.* 1988). The same research group later discovered the two repeating elements, VPE1 and VPE2, after sequencing and aligning the upstream region of five vitellogenins (Spieth *et al.* 1985). The borders of the vitellogenin promoter elements were further refined through comparison to the sequenced vitellogenin promoter containing regions of *C. briggsae* (Zucker-Aprison and Blumenthal 1989). Finally in 1992, it was demonstrated that the VPE1 and VPE2 elements activated the vit-2 promoter (MacMorris *et al.* 1992). First, a vit-2::vit-6 fusion was constructed to assess if 247 bp of the *vit-2* promoter was sufficient to drive normal vitellogenin expression. This construct was then mutated via site-directed mutagenesis and restriction enzyme digestion. The expression patterns of the mutants showed greatly reduced expression. A site-directed mutant with changes in two base pairs of the 3' most VPE1 site eliminated promoter function. From identification of possible regulatory elements to proof of the regulatory effects of these elements entailed seven years of laborious work.

Over the last ten years, many advances should now allow us to greatly reduce the time of such analyses. One advance brought on by the development of automated high-throughput sequencers, is the availability of high quality genomic sequence, with an error rate of less than 10^{-4} (Consortium 1998). Sequence fidelity is a necessity when dealing with the small sequence lengths that constitute regulatory elements where a single incorrect nucleotide can result in an ineffective diagnostic matrix.

A critical component of the approach described in this paper is comparative genomics or phylogenetic footprinting. In this study, the genomic sequence of *C. briggsae* is used to complement the *C. elegans* sequence (Stein *et al.* 2003). *C. briggsae* is physiologically very similar to *C. elegans*, so much so that strains were often confused until Paul Friedman developed a diagnostic technique to differentiate the two species (Friedman *et al.* 1977). Interestingly, the two nematodes are estimated to have speciated 80 to 120 million years ago, greater than human divergence from mouse (Gupta and Sternberg 2003). Of the predicted approximately 20,000 genes in *C. briggsae*, over 12,000 have clear *C. elegans* orthologues and an additional 6,500 have convincing homologs. The high functional conservation, along with the large divergence time of the organisms, makes these two species strong candidates for such comparisons.

Prior to any large scale analysis of the *C. elegans* genome, an algorithm had to be developed. I decided to first develop an algorithm to efficiently predict a single known regulatory motif and then systematically refine the algorithm to become a more general predictor. This approach is unique in the development of regulatory motif discovery; algorithms are generally designed to be general during the entire development. However, this approach provides the researchers with no upper bounds on predictive capacity, and gives no clues to what increases and decreases general prediction power. I chose an approach that would allow me to predict a known regulatory element with the greatest efficiency possible then sequentially refine the algorithm to detect additional elements with only a small reduction in prediction efficiency for those motifs already processed.

The regulatory element chosen as the initial training set was the X box. The X box was first identified as a conserved sequence element in the 5′ promoter containing region of the mouse and human histocompatibility 2 E alpha (H2Eα) genes (Mathis *et al.* 1983). The H2Eα gene is a member of the MHC (major histocompatibility complex) class II family of genes, a group involved in the initiation of the antigen specific immune response. Further analysis of all eight mouse and human MHC class II genes revealed the presence of this 14bp sequence approximately 100bp upstream of all eight genes (Kelly and Trowsdale 1985). The first X box binding factor (RFX1) was isolated using gel retardation assays of X box containing promoter segments. Many additional RFX family members have been since identified, defined by a highly conserved DNA binding domain. In mammals, there are five identified RFX family members. The RFX5 gene results in bare lymphocyte syndrome, the absence of MHC class II gene products in all cell types (Reith and Mach 2001). Mutations in the BBS gene family, which are regulated by the another RFX protein, results in Bardet-Biedl syndrome. Caused by defects in the basal body of ciliated cells, BBS is characterized by many symptoms, mainly retinal dystrophy, polydactyly, mental retardation, and mild obesity (Ansley *et al.* 2003).

The *C. elegans* genome encodes a single RFX family member, DAF-19, which is expressed in the 60 ciliated sensory neurons of the nematode. A loss of function mutation results in the absence of all such cells from the organism. Many *C. elegans* genes have been shown to fall under DAF-19 transcriptional regulation through functional identification, presence of the X box approximately 100 bp upstream of the gene, and expression analysis (Efimenko *et al.* 2005). Also, many of the DAF-19 regulated genes are orthologues of the human BBS genes. Due to the strong evidence and high conservation of the regulatory element, the decision was made to use these genes as the initial test set for OrBS.

The goal of this research is to create a regulatory motif discovery tool to discover common regulatory motifs from sequence sets consisting of co-expressed genes while taking into account information from comparative genomics to bias the program towards biologically meaningful motifs. Starting with the site sampler, we implement the higher-order background and occurrence estimates of Consensus. We further expand the Gibbs sampling approach to include a dynamic motif width algorithm, a scoring function modified to incorporate information derived from comparative genomics, and a unique method for determining when the sampler has converged on the global optimum. OrBS wass initially refined to discover a single well known regulatory element, the X box. The program will later be generalized by sequentially refining it to detect additional known *cis*-acting regulatory elements.

METHODS AND MATERIALS

Hardware and Software

The OrBS program was programmed in C++ and compiled with GNU Compiler Collection (GCC) on SuSE linux 9.1. Gene sequences and location were stored in a local MySQL database to facilitate efficent access. Alignments were created using LAGAN (Brudno *et al.* 2003) with a Perl wrapper.

Data Sets

All *C. elegans* genomic sequence data was acquired from the Wormbase web site, www.wormbase.org, release WS140, date March 26, 2005 (Chen *et al.* 2005). The training set of 14 DAF-19/X box regulated genes was previously used in the identification of BBS3 (Fan *et al.* 2004). The sequence set is consists of 1000 bp upstream of the translation start site of each gene or until the neighbouring gene. The test set of 11 genes was constructed from additional genes with strong DAF-19/X box regulation evidence. The evidence for DAF-19/X box regulation for both gene sets is given in Table 1. *C. elegans* SAGE data and GFP::promoter expression data is available at the *Ce*GEP website, http://elegans.bcgsc.ca (McKay *et al.* 2003).

Table 1 DAF-19/X box Regulated Gene Sets

Gene	Locus	DAF-19 Dependent Expression	X box Dependent Expression	Expression in Ciliated Neurons	References
F33H1.1	*daf-19*	n/a		✓	(Swoboda *et al.* 2000)
Y105E8A.5	*bbs-1*	✓		✓	(Ansley *et al.* 2003; Efimenko *et al.* 2005)
F20D12.3	*bbs-2*	✓	✓	✓	(Ansley *et al.* 2003; Efimenko *et al.* 2005)
Y75B8A.12	*bbs-7*	✓		✓	(Ansley *et al.* 2003; Efimenko *et al.* 2005)
T25F10.5	*bbs-8*	✓		✓	(Ansley *et al.* 2003; Efimenko *et al.* 2005)
F38G1.1	*che-2*	✓	✓	✓	(Fujiwara *et al.* 1999; Swoboda *et al.* 2000)
F59C6.7	*che-13*	✓		✓	(Haycraft *et al.* 2003)
T27B1.1	*osm-1*	✓	✓	✓	(Signor *et al.* 1999; Swoboda *et al.* 2000)
Y41G9A.1	*osm-5*	✓	✓	✓	(Haycraft *et al.* 2001)
R31.3	*osm-6*	✓	✓	✓	(Collet *et al.* 1998; Swoboda *et al.* 2000)
F02D8.3	*xbx-1*	✓	✓	✓	(Efimenko *et al.* 2005; Schafer *et al.* 2003)
F40F9.1a	*xbx-6*	✓		✓	(Efimenko *et al.* 2005)
K08D12.2				✓	(Fan *et al.* 2004)
Y110A7A.20				✓	(Fan *et al.* 2004)
R01H10.6	*bbs-5*	✓		✓	(Li *et al.* 2004)
C27A7.4	*che-11*	✓		✓	(Efimenko *et al.* 2005; Qin *et al.* 2001)
T19A5.4	*nhr-44*	✓		✓	(Efimenko *et al.* 2005)
F53A2.4	*nud-1*	✓		✓	(Dawe *et al.* 2001; Efimenko *et al.* 2005)
Y102E9.1	*odr-4*	✓	✓	✓	(Dwyer *et al.* 1998; Efimenko *et al.* 2005)
F10B5.4	*tub-1*	✓		✓	(Efimenko *et al.* 2005)
D1009.5	*xbx-2*	✓	✓	✓	(Efimenko *et al.* 2005)
M04D8.6	*xbx-3*	✓		✓	(Efimenko *et al.* 2005)
C23H5.3	*xbx-4*	✓		✓	(Efimenko *et al.* 2005)
T24A11.2	*xbx-5*	✓		✓	(Efimenko *et al.* 2005)
R148.1	*xbx-7*	✓		✓	(Efimenko *et al.* 2005)

Algorithm

Basic Algorithm

The algorithm of the Gibbs sampling procedure used by OrBS is an extension of the "basic algorithm" put forth by Lawrence *et al.* (1993), reiterated below. To aid in the understanding of the algorithm, some of the modifications in OrBS are introduced during the initial description of the algorithm but others modifications are described later.

The Gibbs sampling algorithm is essentially a multiple sequence local alignment algorithm which identifies regions of similarity. It is computationally less expensive than traditional local alignment algorithms, such as the Smith-Waterman algorithm (Smith and Waterman 1981), because it examines only a small subset of all possible alignments. It is also quicker than most multiple alignment algorithms, such as CLUSTAL (Thompson *et al.* 1994), because all sequences are aligned simultaneously, not sequentially. Given a set of N sequences, $S_1,S_2,...,S_N$, of length $l_1,l_2,...,l_N$ respectively, each of which contain a similar sequence motif of length W, the algorithm will construct a matrix model θ, a description of a candidate regulatory element. The motif model is maintained in two evolving data structures, the alignment description and pattern description.

The pattern description consists of a W by J matrix, where J is the size of the residue alphabet, four in the case of DNA, and W is the motif width. This position frequency matrix (PFM), $c_{i,j}$, contains the observed frequency of the residue j from 1 to J in position i from 1 to W in the current motif alignment. Such PFM matrices are also used to describe the transcription factor binding site information in both the TRANSFAC (Wingender *et al.* 2001) and JASPAR (Sandelin *et al.* 2004) databases. This description is also stored in the analogous target probability matrix, $q_{i,j}$, defined by

$$q_{i,j} = \frac{c_{i,j} + ps_j}{N + PS}, \qquad (1)$$

where ps_j is the number of pseudocounts for residue j and PS is the sum of all ps_j. Both these descriptions are updated each iteration. In OrBS a pseudocount of the background probability for a given nucleotide was used.

The second evolving data structure is the alignment description. In the Gibbs sampler, the alignment description was described as a vector of motif start positions a_z, for z from 1 to N. In order to ease the expansion of the algorithm, the matrix $a_{z,x}$ was used where x is a valid motif position in sequence z, and $a_{z,x} = 1$ if nucleotide x in sequence z is a start of an element in the alignment, otherwise $a_{z,x} = 0$.

In the initialization stage of the algorithm, one valid motif position is randomly selected from each sequence and added to the motif model (Figure 1b). A sequence is selected randomly, and that sequence's motif is removed from the alignment. The pattern description is updated and all valid motif positions in that sequence are given a weight, creating a probability distribution for the motif (Figure 1c). The higher the weight of a position the more likely that position belongs to the motif model than the background model. A weight of 1 denotes that either case is equally likely. A new motif selected from this distribution and the motif model is updated (Figure 1d). This is iterated for all input sequences and the whole process is iterated many times. At each sampling step, those sequences that resemble the motif model are predominantly chosen, creating a motif model with a stronger consensus. In turn, this new motif model is used to score the valid motif positions in the predictive update step. After many iterations of this process, the algorithm converges on the regulatory element (Figure 1e-f). Upon convergence, a pre-defined number of selection steps occur, refining the motif to the best position in each sequence.

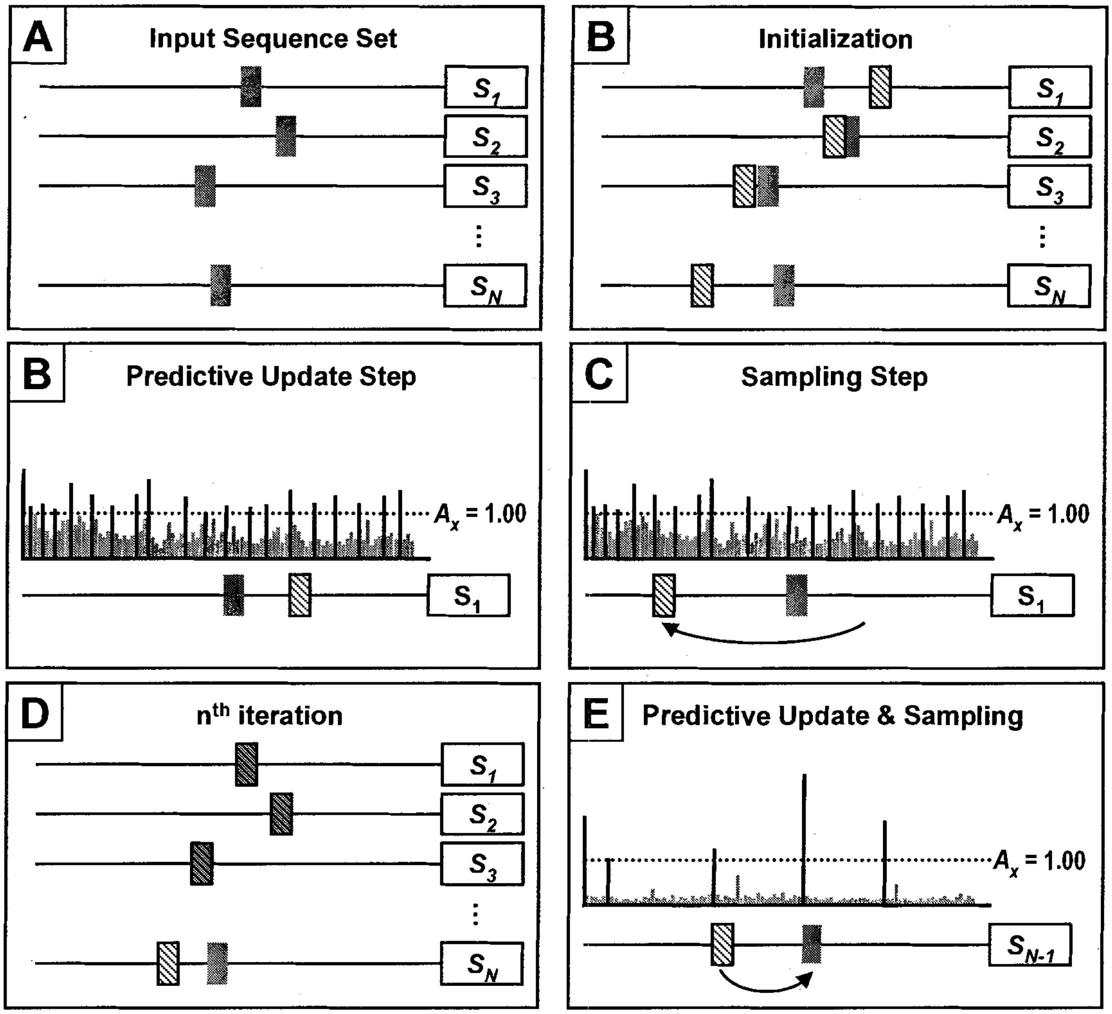


Figure 1 **Basic Gibbs Algorithm**
The genes S1 through SN (white boxes) contain regulatory elements (grey boxes) in the upstream promoter region. The estimated positions of these elements, the motif model selections, are denoted by slash-filled boxes.

Background Model

The background model can be provided by the user in two ways. First, it can be provided by the user in a file. This file must contain a 3rd order model and be in the format where (T|ACG) = 0.45 is represented as ACGT<tab>0.45. Second, if no background probabilities are provided, OrBS will calculate these probabilities from the input sequence. In order to reduce running time, the background probability for each valid motif position is calculated and stored in an array. This prevents repeated calculations for each position during the sampling step. However, if a width change occurs, the probabilities must be recalculated.

Dual Strand Analysis

OrBS examines both strands of the input sequence for motif model inclusion. In order not to double the sequence search space, for any given candidate motif, OrBS will only include the higher scoring strand, thereby keeping the total number of valid positions the same. If an antisense sequence is included in the motif model, the pattern description is updated in the normal manner, however a -1 is stored in the alignment description, as opposed to a 1.

The Scoring Function

The most common scoring function used in Gibbs sampling algorithms for motif detection is the log-likelihood score or LLS (Lawrence *et al.* 1993):

$$LLS(\theta) = \sum_{i=1}^{w} \sum_{j=1}^{J} c_{i,j} \log \frac{q_{i,j}}{b_j}, \qquad \textbf{(2)}$$

where b_j is the frequency of observing residue j in the background model, *Bm*. This is a measure of the pattern's divergence from the background as well as the information content of the pattern. Replacing the frequency term, $c_{i,j}$, with the target probability, $q_{i,j}$, gives us the comparable Kullback-Leiber Information or KLI (Kullback and Leibler 1951). Calculating the probability of all candidate motif positions in a sequence would be very inefficient because the pattern description would require updating for each position examined. Instead a simpler equation that is proportional to the scoring function is used. In this case of the LLS and the KLI, this weight score is:

$$W_z(x) = \frac{P(x|\theta)}{P(x|Bm)} = \prod_{i=1}^{w} \frac{q_{i,x_i}}{b_{x_i}}, \qquad (3)$$

where the weight assigned candidate subsequence starting at x from sequence z, $W_z(x)$, is the probability the sequence exists in the motif model divided by the probability the sequence exists in the background model (Lawrence *et al.* 1993). From this weight distribution, a sequence is sampled and added to the alignment from which θ is generated.

One problem with using the KLI as described by Lawrence *et al.* (1993) is the effect of the motif width on the score. Given a motif model of a defined width, θ_w, generated from the alignment matrix a, the resulting KLI of that model will always be less than a model defined by the same alignment matrix but with a greater width. In a more mathematical form, the statement KLI(θ_w) < KLI(θ_{w+1}) for a given a, is always true. In the original Gibbs this does not pose a problem, but in the dynamic width algorithm implemented in OrBS, the algorithm would tend towards a motif of the smallest valid width. A simple, naïve, solution is to normalize the score by the width of the motif. However, this will indirectly cause the selection of shorter motifs due to the greater likelihood of high conservation among shorter sequences, stochastically. In order to circumvent this issue, a zero normalized information content score is instead used during the dynamic width and phase shift procedure. This normalization is made by subtracting an expected KLI score from the each column KLI of the motif. The expected score of 0.25 was determined by the median IC score for the outermost column of all motif matrices in the TRANSFAC database. There is no reason that this scoring function must only be based on the distribution of residue frequency. In OrBS, the scoring function is modified to include sequence conservation between species as well. In order not to limit the methodology of determining of orthology contribution, the orthology bonus is kept simple:

$$OrBS(\theta) = \left(1 + \frac{\sum_{i=1}^{w} h_i}{w}\right) \sum_{i=1}^{w} \sum_{j=1}^{J} q_{i,j} \log \frac{q_{i,j}}{b_j}, \tag{4}$$

where h_i is the cross-species conservation bonus of position i normalized to 1. It rewards cross-species sequence conservation of the sequence. It is commonly held that there is a correlation between regions of high conservation between species and sequence functionality. This bonus biases the motif alignment towards regions of conservation. The resulting weight function is:

$$W_z(x) = \left(1 + \frac{\sum_{i=1}^{w} H_{z,x}}{w}\right) \prod_{i=1}^{w} \frac{q_{i,x_i}}{b_{x_i}}, \tag{5}$$

where $H_{z,x}$ is the orthology bonus assigned to residue x in sequence z normalized to 1.

Dynamic Width and Phase Shifts

Lawrence *et al.* (1993) describe a possible solution to the issue of phase shifts. A phase shifted result is a reported solution that is shifted a couple base pairs to the right or left of the optimal solution. A phase shifted solution results from the biased sampling of similar motifs. If a shifted optimal solution motif is selected early on in the algorithm, when sampling biases are small, it will likely perpetuate throughout the execution of the program. They suggest that after a given number of iterations, the algorithm should examine the score of the motif models resulting from a shift in all alignment starting positions left or right to a maximum shift distance. I expand this proposed modification to also examine those motif models that vary in width, *W*, a set number of nucleotides from the current model. Thus, this will not only correct any phase shift anomalies that occur during the operation of the algorithm but any incorrect assumptions of motif size.

The procedure is basic and simply creates a temporary *q* matrix with the width determined by the maximum valid phase and width shift. Then all valid alternative motif models that fall within this temporary *q* are scored. A new motif model is then selected by sampling from the distribution of these scores.

Motif Occurrence

The motif occurrence problem is addressed in OrBS with the algorithm utilised in Motif Sampler (Thijs *et al.* 2002). This algorithm is computationally intensive and requires the inclusion of an additional step. During the expected occurrence step, the probability of observing *o* occurrences of the current motif in each sequence is calculated independently for *o* = 0, to a given maximum $O_{\max}$, using the formula:

$$\gamma_z(o) = P(Q_z = o \mid S_k, \theta, Bm), \tag{6}$$

where Q_z is the actual number of occurrences of the motif θ in sequence S_z. Thjis *et al.* (2002) show how equation 6 can be further expanded by applying Bayes' theorem:

$$\gamma_z(o) = \frac{P(S_z \mid Q_z = o, \theta, Bm) P(Q_z = o \mid \theta, Bm)}{P(S_k \mid \theta, Bm)}, \tag{7}$$

where the numerator is the probability that the sequence is generated from the background model and *o* occurrences of the motif θ multiplied by the probability that the actual number of occurrences is *o* given the motif and the background model. The denominator is essentially the sum of all possible number of occurrences, resulting in the final equation:

$$\gamma_z(o) = \frac{P(S_z \mid Q_z = o, \theta, Bm) P(Q_z = o \mid \theta, Bm)}{\sum_{m=0}^{\infty} P(S_z \mid Q_z = o, \theta, Bm) P(Q_z = o \mid \theta, Bm)}, \tag{8}$$

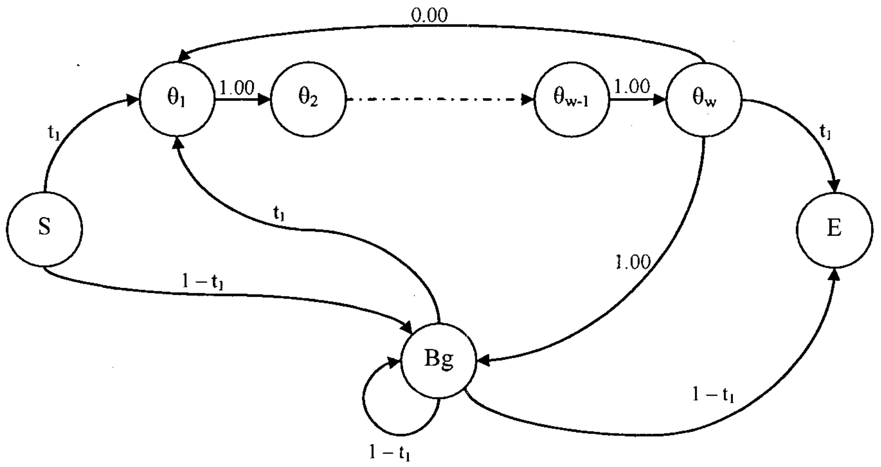
where $P(S_z|Q_z = o, \theta, Bm)$ can be solved in linear time using a modified form of the forward algorithm. This involves the adaptation of the motif and background model into a single HMM where $P(Q_z = o)$ is the *a prior* of finding the motif model o times (Thijs *et al.* 2001). This equation is calculated for each c up to a O_{max}. The resulting values allow for the calculation of the expected occurrence of the motif as follows:

$$E(Q_z) = \sum_{o=0}^{O_{\max}} o\gamma_z(o). \tag{9}$$

In order to reduce computational time, in the implementation of OrBS, O_{max} has been hard coded to a value of 4. However, this may be changed to a user definable parameter in a later version. Even with the time saved using the Forward Algorithm, this is the slowest step of the algorithm. For that reason, the period of the expected occurrence step is user definable (parameter Oi). A graphical implantation of the Forward Algorithm can be seen in Figure 2. While the structure and the emission probabilities of the Markov chain remains the same, the transition probabilities are dependent on the value of o.

Occurrence HMM Probabilities:
L = length of sequence
W = width of motif

Option 1 (occ = 1): $t_1 = \dfrac{1}{L - w + 1}$



Option 2 (occ = o):
positions: p = L – o(w-1)
configurations: Choose(p,o)

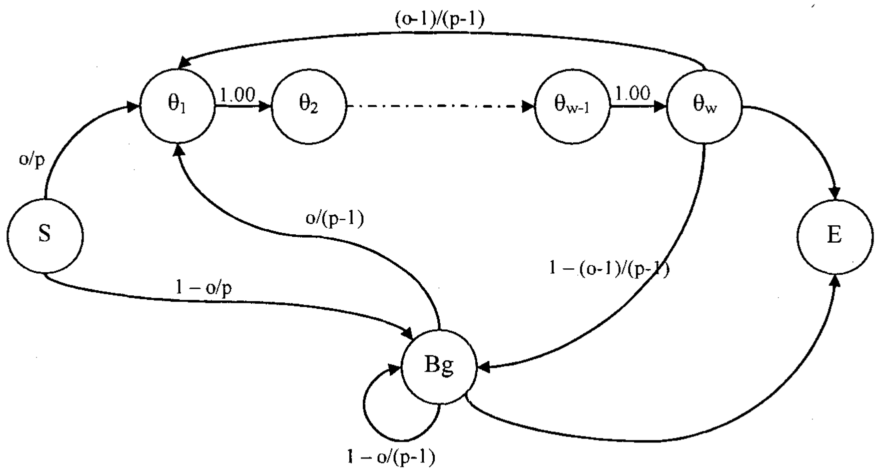


Figure 2 **Structure of Markov chain in expected motif occurrence algorithm**
The arrows of the represent possible paths into and out of each state. The transmission probabilities are along the arrows paths. The emission probabilities of each state are determined by the motif model (*θ*) and background model (*Bg*) used to create the HMM. The state S is the starting state and the state E is the ending state. These represent imaginary nucleotides bordering the first and last nucleotides, respectively.

Motif Detection

A motif is reported by OrBS when a user defined number of parallel running Markov chains converge upon a similar motif pattern description matrix. The similarity between these two matrices is a Pearson correlation coefficient (PCC) (Schones *et al.* 2005). Given two motif models, θ_X and θ_Y, the PCC of the two respective columns, x and y:

$$PCC(\theta_{X,x},\theta_{Y,y}) = \frac{\sum_{j=0}^{J}\left(c_{Y,y,j} - \overline{c_{Y,y}}\right)\left(c_{X,x,j} - \overline{c_{X,x}}\right)}{\sqrt{\sum_{j=0}^{J}\left(c_{Y,y,j} - \overline{c_{Y,y}}\right)^2 \sum_{j=0}^{J}\left(c_{X,x,j} - \overline{c_{X,x}}\right)^2}}, \qquad \textbf{(10)}$$

where a score of -1 is perfect inverse correlation, 0 is no correlation, and 1 is perfect correlation. The PCC is a measure of linear correlation and therefore cannot identify shift variants of similar motifs. However, because the calculation of the PCC is simple and computationally inexpensive, the PCC is calculated for all alignments of the two matrices, reporting only the greatest score.

If the two matrices have a PPC greater than the user defined threshold, the Gibbs sampling algorithm is determined to have converged. The motif model with the greater OrBS score is refined through several selection steps and returned as an observed motif. The values of the alignment description are then transferred to the masking matrix, *m*. If $m_{i,j}$ holds a value other than 0, that position cannot be included in the motif model. The algorithm then attempts to identify additional motifs in the sequence set. The OrBS algorithm terminates when a user defined number of iterations occurs between convergence events.

Parameter Refinement

The initial training consists of the upstream region of 14 genes previously used to create an X box consensus pattern (Table 1) (Fan *et al.* 2004). The X box is a well conserved, 14 nucleotide motif, typically approximately 100bp upstream of the affect genes start site. In *C. elegans*, RFX is expressed specifically in ciliated neurons. The X box set, unless otherwise stated, comprises of the upstream region, up to 1 Kbp or the distance to the neighbouring upstream gene. The X box set can be regarded as an easy TFBS to identify due to the large size and tight consensus of the sequence. No other TFBSs are known to be shared among these sequences, and it is currently assumed there is none due the restricted common expression pattern.

Test Statistics

Three common statistics were used to measure the effectiveness of OrBS in identifying the regulatory elements. The first is sensitivity (*Sn*), a measure of how well the algorithm identifies actual occurrences of the motif and is defined as

$$Sn = TP/(TP+FN),$$

where *TP* is the number of true positive sites and *FN* is the number of false negative sites. The definition of a TP is a prediction that encompasses at least 80% of the known regulatory element. The second is the Positive Predictive Value (*PPV*) which gives the fraction of positive observations that are true:

$$PPV = TP/(TP+FP),$$

where *FP* is the number of false positive sites. The last statistic, Specificity (*Sp*) is a measure of how well the algorithm “ignores” incorrect sites:

$$Sp = TN/(TN+FP),$$

where *TN* is the number of true negative sites. These statistics were used in combination with the OrBS score to determine the best parameter set.

Convergence and Termination

The canonical X box set was used to test and optimise the convergence parameters of OrBS with an alternate number of parallel Markov chains (-M) of 2, 5 or 10, and an alternate number of similar motifs required for a convergence event (-CN) of 2 or 3. Also, two different Pearson correlation coefficient thresholds were tested, 0.75 and 0.90. The *a priori* of motif occurrence will be refined in a later test and since it was known that all sequences contain the X box motif, no occurrence estimation step is performed. In order to ensure all motif alignments that would converge within an acceptable timeframe did converge, the maximum number of iterations between convergence events (-Ci) was overestimated with a value of 10,000.

Occurrence and Noise

After the convergence parameters were selected, the X box set was also used to test the effect of noise, or the inclusion of non-coregulated genes, on the OrBS algorithm. Three additional test groups were created. Each set includes a defined number of upstream regions randomly selected from *C. elegans* genes in addition to the canonical X box set (Appendix B). These sets were also used to test three values of the occurrence *a priori* estimates, 0.25, 0.50, and 0.75. More than one test set was constructed for each noise level in order to reduce any set specific effects on occurrence estimation. For this comparative analysis, each sequence was examined as a single site, where a positive was an observation in that sequence and a negative as no observation. Only those parameter combinations that were deemed acceptable in the previous tests were tested.

Orthology

In all previous tests, no orthology bonuses were assigned to any of the sequences.

In order to test the effects of the new scoring function, the X box test sets used in Occurrence and Noise optimization tests were reanalysed with the assigned orthology bonuses. The orthology bonus used was based on a LAGAN (Brudno *et al.* 2003) alignment with the upstream region of the orthologous *C. briggsae* gene. The orthologues were previously assigned by Lincoln Stein using a best reciprocal BLASTP procedure and are available at ftp://ftp.wormbase.org/pub/briggsae/orthologues_and_orphans/orthologues.txt.gz (Stein *et al.* 2003). The orthology bonus assigned to each residue was 1 if the nucleotide was conserved in the pairwise alignment and 0 if it was not.

Performance on Known Sets

X box Regulated Genes

The previously predicted X box motif with the highest width normalized OrBS score was used to scan the *C. elegans* genome for additional occurrences. This was implemented using the TFBS module available for Perl (Lenhard and Wasserman 2002). The scores assigned to the original 14 X box gene promoters were used to determine an occurrence cut-off score of 17.18, the lowest score of the set. The upstream 1000 bps or distance to the neighbouring gene for all *C. elegans* genes in the WS140 release was examined. Genes whose promoters contained putative X box with scores greater than the threshold were assessed for additional evidence of cilia specific expression. Both promoter::GFP fusion expression profiles and tissue specific SAGE data available from the *Ce*GEP website (McKay *et al.* 2003). The SAGE library for FACS sorted ciliated neurons was compared to SAGE library for both FACS sorted pan-neural cells and FACS sorted muscle cells. The tag counts for each transcript was compared using the Chisquare test (Kal *et al.* 1999). Expression profile categories from the promoter::GFP fusion data are quite broad. Out of the 40 expression categories, 6 include cilated neurons: amphids, phasmids, head neurons, tail neurons, and mechanosensory neurons.

Standardized Test Set

Each of the 156 data sets in standardized test set was downloaded from http://bio.cs.washington.edu/assessment/ and run with same settings as for the previous test sets. In the case that more than one motif is predicted for a given dataset, the motif with the greatest OrBS score was reported. The resulting output was then converted to agree with the format required by the Assessment of Computational Motif Discovery Tools submission form and uploaded to the system. The results were returned as an excel datasheet (Appendix D).

Additional *C. elegans* Datasets

In addition to the X box set used in the parameter optimization of OrBS, two alternative *C. elegans* test sets were used to test the efficiency of the algorithm to detect transcription factor binding sites, two of which were used in the analysis of a similar program CompareProspector (Liu *et al.* 2004). The upstream region for each gene was determined by taking the sequence from the starting codon of the gene to the closer of either the neighbouring upstream gene or 1 Kbp.

In a recent paper describing another TFBS discovery program, two *C. elegans* data sets are described (Liu *et al.* 2004). The first is a set of genes identified by microarray analysis for increased embryonic expression in *par-1* mutants as compared to *skn-1* mutants (Gaudet and Mango 2002). These two strains produce excess or no pharyngeal cells, respectively. This method identified 240 genes with an average 2-fold increase in expression in the *par-1* mutants. Previous studies have successfully identified the PHA-4 binding site using the upstream sequences of these genes (Liu *et al.* 2004). Of the 240 genes identified, 199 had at least 150bp separating the start codon with the neighbouring upstream gene, as described by Wormbase, and 148 of these have been assigned *C. briggsae* orthologues. The second set consists of three genes, *mec-3*, *mec-4*, and *mec-7*, all of which have been shown to contain the UNC-86::MEC-3 heteroligomer binding sites (Duggan *et al.* 1998; Xue *et al.* 1992; Xue *et al.* 1993). Both *unc-86* and *mec-3* are involved in the differentiation of touch receptors in *C. elegans*. The UNC-86::MEC-3 binding sites occur multiple times in each promoter region they affect. This was used to test the effectiveness of the occurrence algorithm implemented in OrBS to detect multiple occurrences (Duggan *et al.* 1998).

RESULTS

Algorithm

In recent years there have been a number of computational approaches to the discovery of transcription factor binding sites. The majority identify a motif or sequence common to the input sequences with some allowance for mismatches. The two most common methods are the heuristic alignment approach and the word counting approach. The latter is far more computational intensive, examining the frequency of all possible sequences of a given size. In this study, we decided to implement the former technique, specifically an expansion on the Gibbs sampling algorithm first described by Lawrence *et al.* (1993).

A Gibbs sampler is in a multiple local alignment tool, identifying repeated motifs in the given input sequence. A more detailed explanation of the original Gibbs sampler can be found in the Methods and Materials section of the thesis. Once a working instance of the Gibbs Sampler was coded in C++, the algorithm was sequentially modified to include the additional features outlined below.

Background Model

The first major modification to the original algorithm was the use of a species specific higher order background model. Such models have previously been shown to increase the predictive power of the Gibbs sampling algorithm in prokaryotes and

Arabidopsis (Marchal *et al.* 2003; Thijs *et al.* 2001). If no background probability file is provided, OrBS will compute a background model from the input sequences. The order of the background model is based on the size of the input sequence set with a maximum order of 3.

Scoring Function

All Monte Carlo Markov Chain methodologies require a function to score the goodness of the current state of the chain. The most common and basic scoring functions take into account both the conservation of the motif and the divergence of the motif model from the background model. However, such functions do not utilize all the useful biological information available, such as distribution of positive selective pressure on the input sequences. This information is available in the form of interspecies sequence conservation and can be estimated through sequence alignment. In OrBS, the algorithm maximizes the typical Kullback-Leiber Information (KLI), a log-likelihood variant. However, the score is then multiplied by the average conservation of the candidate subsequence. The conservation is based only on the nucleotides to be included in the motif alignment and not an average over a larger window as used in some algorithms (Liu *et al.* 2004). This biases the algorithm towards those subsequences which show evidence for positive selection while maintaining the sequence space. The nucleotide orthology bonus is provided by user in the sequence input file. If no orthology bonus is available for a given sequence, or the entire sequence set, the bonus is assumed to be 0 and the algorithm proceeds using the original KLI score.

Dynamic Widths and Phase Shifts

Lawrence *et al.* (1993) address a couple of concerns or 'defects' with the basic algorithm they originally put forth. Two issues are the phase shift and static width problems. The width problem is the easier to describe. In the site sampler, and most of the *de novo* motif prediction programs examined, the width of the unknown motif is fixed and defined by the program's user. This is a contradiction in the sense that such programs assume the user has knowledge of the unknown motif. While an educated guess can be made based on known motif sizes, the user cannot and should not be required to know the motif width.

The second issue is what Lawrence refers to as the phase problem. The phase problem is defined as the shortcoming of the algorithm to become stuck in a solution that is a shifted form of the optimal solution (Lawrence *et al.* 1993). For example, assume there exist five sequences, $S_1,...,S_5$ which share a common motif. Let the optimal solution for this motif be the alignment positions 20, 14, 3, 37 and 28 for the five sequences respectively. However, if the algorithm were to initially choose the positions 19 and 36 for the sequences S_1 and S_4, it is unlikely the optimal solution will be reached. Instead a phase shift form of the optimal solution defined by the alignment positions 19, 13, 2, 36 and 27 is the more probable outcome.

Lawrence *et al.* (1993) propose a possible solution to the phase shift problem by the insertion of an additional step. This step involves the comparison of the current motif alignment description to alternative alignment descriptions that are shifted right or left up to a defined number of residues. A new motif is sampled, in an algorithm analogous to the main sampling step, using weights defined by the scoring function. This solution can be further expanded to the static width problem. In OrBS, not only are the phase shift variants of the current motif considered but also larger and smaller width variants. This allows the width of the motif to be dynamically updated throughout the prediction process and reduces the negative effects of a poorly chosen width, an attribute missing from most of the current motif prediction algorithms.

In OrBS, the width and phase shift problems are addressed with a single algorithm. Following the suggestions of Lawrence *et al.* (1993), after a specified number of sampling iterations (-Wi), a window is drawn around the current alignment of which the size is defined by the maximum allowed width change (-WS) and phase shift (-PS). However, if each of the resulting alignments was scored using the previously defined OrBS score, larger motifs would be selected preferentially. By assigning width independent scores to each candidate alignment, a normalized Kullback-Leiber Information (nKLI) score was developed:

$$nKLI = \sum_{i=0}^{W} \sum_{j=0}^{J} \left(q_{i,j} \log_2 \left(q_{i,j}\right) - es\right), \tag{11}$$

where *es* is the expected KLI score for a regulatory motif. The value of *es*, 0.25, is the median KLI score for the outermost column of all motif matrices in the TRANSFAC database. This is far below the overall median column KLI of 0.87, allowing for the extension of weak motifs but still higher from the median KLI column score or 0.08 for a random alignment of *C. elegans* upstream sequence.

Motif Occurrence

Co-expression is not co-regulation, and thus it is unlikely that all upstream regions of a set of co-expressed genes are controlled by a single transcription factor and contain the associated regulatory element. For this reason it is necessary for an effective motif prediction algorithm to allow for the absence of the current motif in any given sequence of the sequence set. Along the same lines, it can be helpful if the algorithm allows for multiple occurrences of the motif in a single sequence, leading to a truer and more robust diagnostic motif pattern. Although difficult to implement, the methodology described originally by Thijs *et al.* (2002) seems the most biologically correct. Using the current motif model and the background model, the OrBS algorithm predicts the expected occurrence of the motif model.

Motif Detection

One of the major challenges of Markov Chain Monte Carlo methods is determining when convergence has occurred or, more simply, the optimal solution has been observed. Convergence is said to have occurred when the Markov chain reaches equilibrium (Cowles and Carlin 1996). A defining feature of the equilibrium state is that it is independent of the starting state, thus every Markov chain for a given Gibbs sampling algorithm will, over some number of iterations, eventually reach this equilibrium state. However, determining the minimum number of iterations required to reach the equilibrium state is a topic of much discussion and typically involves complex mathematical inference (Cowles and Carlin 1996). In OrBS, we attempt a novel approach in which attempt to determine convergence by observing Markov chain state similarity above a given threshold.

OrBS is also able to detect multiple motifs in the provided sequence set. This is accomplished sequentially, through the same methodology as used in many previous algorithms (Aerts *et al.* 2003b; Frith *et al.* 2004; Thijs *et al.* 2002). Once the algorithm has converged on a motif, the sub-sequences that make up the alignment for that motif are masked. The algorithm then continues sampling from the updated sequence set.

The iterative Gibbs sampling procedure of OrBS is terminated when it is deemed no further patterns can be identified. This is defined by the passing of a given number of iterative steps in which convergence has not been observed. Convergence is observed when two parallel running Markov chains exist in states that show a similarity above a given threshold at the same time point. Similarity is determined though the PCC of the pattern description matrices of the two motif models. See the Methods and Materials section for a more detailed explanation of the convergence determination procedure.

There exist five variables relating to convergence and sampling termination that are user definable: the number of parallel chains (-M), the degree of correlation required for convergence (-CS), the frequency of the convergence check (-Ci), the number of correlated chains required for convergence (-CN), and the number of iterations without a convergence event causing termination (-Mi). When two or more motifs models are identified to have a high correlation coefficient, the higher scoring motif model is then selected, refined through several selection steps and returned as an observed motif. The sites of the observed motifs are masked and the sampling iterations continue.

Program Refinement

OrBS is implemented in a single, stand-alone, executable file with a simple command line interface. Execution of the program without passing any parameters to it results in a listing and explanation of all available parameters:

```
OrBS Options:
 -s  sequence file in proper format (mandatory)
 -B  file containing the a background model
 -W  motif width the gibbs sampler is initialized with (default 10)
 -WS maximum motif width shift (default 2)
 -PS maximum motif phase shift (default 2)
 -Wi iterations between motif shift algorithm (default 2)
 -Si number of selection steps after sampling iterations (default 3)
 -M  number of motifs to identify(default 10)
```

```
-H  homology bonus (default 1.00)
-CN number of similar motifs required for convergence (default 2)
-CS minimum similarity score for convergence (default 0.80)
-Ci convergence update interval (default 2)
-Oi occurrence update interval (default 1)
-v  verbose mode
```

Only the file containing the set of genes expected to share transcriptional regulation is required. The sequence file contains a list of sequences, one per line, with the line header 'SEQUENCE:' followed by the optional orthology score list where the scores are comma separated and the line header of 'ORTHOLOGY:'. If a background nucleotide frequency model is provided, it must be a 3rd order model with a line format of:

nucleotide(s)<tab>frequency<line return>,

where nucleotides is in the format of the last residue given the preceding nucleotides. The algorithm will report an error if either of these two files are formatted incorrectly.

Convergence and Termination

The results of the convergence parameter optimisation are an average of three replicate runs of OrBS with the previously defined settings are shown in Table 2. For the X box training set, all sets of parameters we able to identify the X box except for one; (M, CN, CS) = (5, 3, 0.90). For this set of parameters the parallel motif chains never converged and no motifs were reported for two of the replicate runs. As expected, all parameter sets with a PCC threshold of 0.75 predicted many more motifs than their more stringent alternative. In the results from the parameter set with a 0.90 PCC threshold, no other motif patterns were common to all replicates from that set.

Table 2 Determination of efficient convergence parameters.

Parallel Motif Chains (-M)	2		5				10			
Motifs Converged (-CN)	**2**		**2**		**3**		**2**		**3**	
PCC threshold (-CS)	**0.75**	**0.90**	**0.75**	**0.90**	**0.75**	**0.90**	**0.75**	**0.90**	**0.75**	**0.90**
Motifs Predicted	10.00 ±7.1	1.00 ±0.0	20.33 ±1.7	3.33 ±1.2	1.67 ±0.9	0.33 ±0.0	18.33 ±3.8	14.33 ±8.2	3.67 ±0.9	1.00 ±0.0
Average Raw Score of Motifs	10.55 ±3.2	18.03 ±0.3	9.73 ±2.5	12.71 ±4.3	15.57 ±4.5	20.26 ±0.0	9.73 ±2.4	9.89 ±2.8	13.02 ±4.0	19.06 ±0.9
Average Number of Iterations Between Convergence	2174 ±2012	1017 ±267	408 ±480	2798 ±2752	2782 ±3213	1428 ±0	113 ±122	1337 ±1725	1362 ±1142	1246 ±904
X box Predicted	yes	yes	yes	yes	yes	no[‡]	yes	yes	yes	yes
Normalized X box Score	1.15 ±0.1	1.15 ±0	1.13 ±0.1	1.05 ±0.0	1.16 ±0.1	1.07 ±0.0	1.15 ±0.1	1.17 ±0.1	1.15 ±0.1	1.18 ±0.1
Number of Iterations to Converge on the X box motif	89 ±75	1017 ±267	9 ±3	93 ±83	240 ±167	1428 ±0	7 ±0	65 ±46	58 ±31	1246 ±904

[‡] the X box motif was identified in some but not all of the runs

Two sets were deemed acceptable for further analysis. These were the parameter combination A (5, 2, 0.90) and combination B (10, 2, 0.90). While both of the other sets (2, 2, 0.90) and (10, 3, 0.90) scored well, the small number of observed motifs may have posed a problem during later stages of the algorithm's development.

Occurrence and Noise

When the motif occurrence estimation step was incorporated into OrBS, the X box motif was remained easily identified (Table 3). Only the results for the *a priori* occurrence estimate of 0.50 are presented. The results for the 0.25 and 0.75 estimated produced primarily empty motifs and motifs with observations in all sequences respectively, and are of little interest (data not shown). The X box motif was identified in all test sets containing up to a sequence noise value of 50 percent. Parameter combination A tends to perform slightly better than combination B at lower noise levels in average normalized motif score and in all test statistics for the identification of the X box motif. However, B did perform better at the 50% noise level as well as an identification of the X box motif in one of the four test sets with a noise percentage of 67. Neither parameter combination identified the X box in any of the test set with 75% noise.

Table 3 **Effectiveness of occurrence algorithm in removing noise.**

Noise (% sequences)	0%				25%				50%				67%				75%			
Configuration	A		B		A		B		A		B		A		B		A		B	
Motifs Predicted	3.00	±1.4	2.50	±1.1	3.25	±1.9	2.75	±1.8	4.50	±2.1	3.50	±1.1	2.00	±2.1	2.25	±1.5	2.75	±1.1	4.50	±2.2
Average Raw Score of Motifs	13.21	±4.3	14.86	±3.3	13.98	±3.3	13.65	±4.3	13.49	±4.4	13.16	±4.7	13.56	±4.5	14.00	±5.1	10.09	±0.9	10.34	±2.9
Average Number of Iterations Between Convergence	4114	±3684	813	±923	2673	±2458	465	±393	3731	±3201	841	±801	2768	±2360	961	±783	2566	±2422	536	±504
X box Predicted	yes		yes		yes		yes		yes[†]		yes[†]		no		no[‡]		no		no	
Normalized X box Score	1.13	±0.1	1.13	±0.1	1.19	±0.1	1.10	±0.1	1.17	±0.1	1.22	±0.1			1.12					
Sensitivity	0.95	±0.1	0.93	±0.0	0.95	±0.1	0.88	±0.1	0.93	±0.1	0.96	±0.1			0.86					
Positive Predictive Value					0.96	±0.1	0.96	±0.1	0.88	±0.0	0.92	±0.0			0.63					
Specificity					0.93	±0.1	0.93	±0.1	0.88	±0.0	0.91	±0.0			0.75					
Number of Iterations to Converge on the X box motif	467	±194	49	±42	1433	±1259	296	±291	856	±588	214	±335			1260					

[†] the X box motif was not the highest scoring motif identified

[‡] the X box motif was identified in some but not all of the four runs

Orthology

The previous data set was reanalysed with the parameter combination B, however, this time an orthology bonus was given to each nucleotide as described above (see Methods and Materials). The inclusion of comparative sequence information resulted in prediction of the X box motif in higher noise percentage data sets (Table 4). The performance, based on the test statistics, between OrBS with and without the orthology bonus is comparable at low noise levels. Only by the inclusion of the orthology bonus to the scoring function was OrBS able to consistently identify the X box at a noise value of 67 percent, but not any higher. The average number of iterations required to identify the X box was reduced for low noise levels but increased when more noise was present. The total number of identified motifs increased when using comparative sequence analysis, resulting in an increase of total run time. A graphical summery of the statistical diagnostic score for parameter set A, B and B with the orthology bias can be seen in.

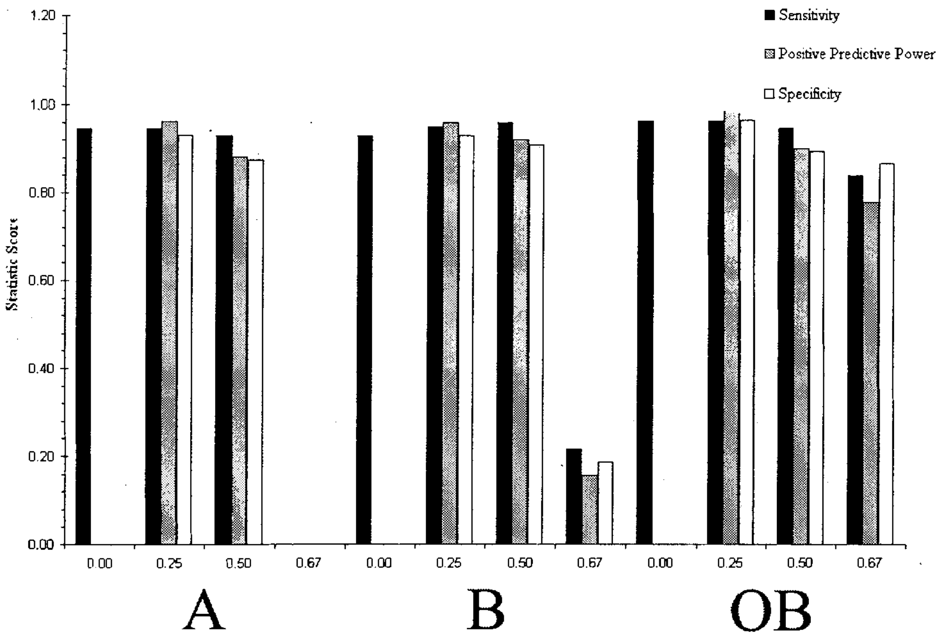


Figure 3 **Graphical comparison of different parameter sets**

Table 4 Orthology bonus

Noise (% sequences)	0%				25%				50%				67%				75%			
Configuration	**B**		**OB**		**B**		**OB**		**B**		**OB**		**B**		**OB**		**B**		**OB**	
Motifs Predicted	2.50	±1.1	3.00	±0.0	2.75	±1.8	3.25	±1.1	3.50	±1.1	4.50	±1.7	2.25	±1.5	6.25	±1.3	4.50	±2.2	2.50	±1.8
Average Raw Score of Motifs	14.86	±3.3	14.31	±3.7	13.65	±4.3	13.33	±4.2	13.16	±4.7	13.09	±4.5	14.00	±5.1	12.21	±3.6	10.34	±2.9	10.56	±1.2
Average Number of Iterations Between Convergence	813	±923	401	±338	465	±393	528	±559	841	±801	709	±700	961	±783	892	±594	536	±504	891.1	±805
Average Orthology Bonus			0.50	±0.0			0.42	±0.1			0.38	±0.1			0.31	±0.1			0.34	±0.1
X box Predicted	yes		yes		yes		yes		yes[†]		yes[†]		no[‡]		yes[†]		no		no	
Normalized X box Score	1.13	±0.1	1.14	±0.1	1.19	±0.1	1.16	±0.1	1.22	±0.1	1.16	±0.1	1.12		1.09	±0.1				
X box Orthology Bonus			0.52	±0.0			0.52	±0.0			0.50	±0.0			0.45	±0.0				
Sensitivity	0.93	±0.0	0.96	±0.0	0.95	±0.1	0.96	±0.1	0.96	±0.1	0.95	±0.0	0.86		0.84	±0.1				
Positive Predictive Value					0.96	±0.1	0.98	±0.0	0.92	±0.0	0.90	±0.1	0.63		0.78	±0.2				
Specificity					0.93	±0.1	0.96	±0.1	0.91	±0.0	0.89	±0.1	0.75		0.87	±0.1				
Number of Iterations to Converge on the X box motif	49	±42	28	±24	296	±291	210	±130	214	±335	303	±212	1260		1983	±1798				

[†] the X box motif was not always the highest scoring motif identified

[‡] the X box motif was identified in only one the four runs

Performance on Known Sets

X box Controlled Genes

Two additional tests were performed in order to test the effectiveness of OrBS in the detection of the X box element. First, the program was run on a test set of 11 genes with strong evidence of X box regulation (Table 1), to ensure the algorithm was over fitted to the training set. Using the run parameters determined from the previous refinement steps, OrBS was able to detect the X box motif in all four test runs with an average normalized OrBS score of 1.24, a sensitivity of 0.89.

An effective motif detection algorithm not only needs to recognize patterns in the input sequence but also output that pattern in a format that can be used as a diagnostic for additional occurrences of that motif. As previously stated, it is unwise to assume that all co-expressed genes are co-regulated. It is equally unwise to assume all co-regulated genes will be detected by a given co-expression assay. For this reason, it is important to confirm the motif OrBS outputs is a good diagnostic. We took the predicted X box motif with the greatest per nucleotide OrBS score and used it to scan the genome for additional X box regulated genes. A sequence logo is a standard and useful way to graphically examine sequence motifs (Schneider and Stephens 1990). The sequence logo of the OrBS predicted X box consensus and the X box consensus used by Fan *et al.* (2004) for the same data set are expectedly very similar (Figure 4). The observable differences are the additional leading nucleotide of the OrBS prediction and slight variances in nucleotide frequency and information content of the shared residues. These variances are due to the palindromic nature of the X box regulatory element and that OrBS allows for motif occurrence on both strands, where all of the motif occurrences in the Fan *et al.*

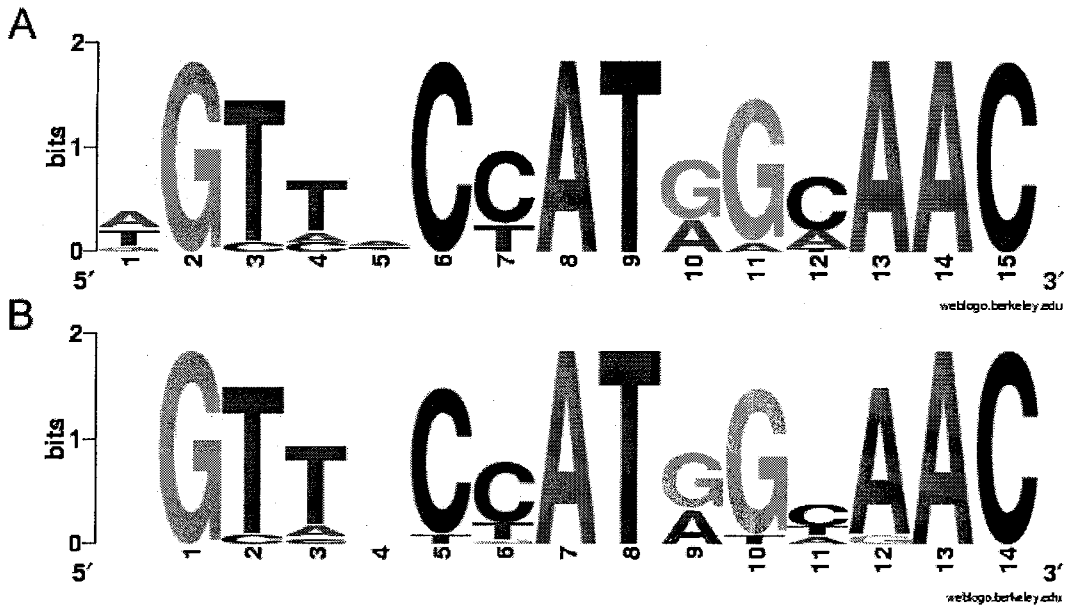


Figure 4 Sequence logo comparison of the X box consensus produced by OrBS and Fan *et al.* (2004)
Both sequence logos were created using WebLogo (http://weblogo.berkeley.edu) (Crooks *et al.* 2004) and the X box training set. (A) This sequence logo was created from the sequence alignment of the highest scoring OrBS prediction of the X box regulatory element. (B) The sequence logo of the hand-curated consensus used to scan the genome for additional X box sites by Fan *et al.* 2004.

(2004) alignment are on the sense strand.

Using the Perl module TFBS and the positional frequency matrix created from the OrBS predicted motif we identified 42 additional possible DAF-19 regulated genes containing an X box sites within 1 Kb of the translation start site (Table 5). In order to validate these predictions, expression data from CeGEP and functional annotation from Wormbase was examined. Of the 30 genes with SAGE tags, 14 show ciliated neuron expression and 4 show over-expression in ciliated cells (Figure 5). All 5 genes identified with available promoter::GFP expression data are expressed in tissues known to contain ciliated cells. The upstream region of the orthologous *C. briggsae* genes were scanned for the occurrence of X box elements. Using an arbitrarily lower cut-off score of 13, to allow for species differentiation from the consensus, X box elements were discovered in the upstream regions of 21 of the 32 orthologous genes with 1 Kb of the translation start site.

The highest scoring X box site is found 50bp upstream of the recently annotated gene F58A4.14. The gene product of F58A4.14 is the orthologue of the human protein BBS4 (BLAST e-value = 2E-56). However, it is puzzling that this gene has not been previously identified since BLASTing the human BBS4 protein against a translated *C. elegans* genome (tBLASTn) identifies the region of F58A4.14 fairly tightly with 9 of the 10 HSPs occurring within the current gene models boundaries (Figure 6).

The genes Y37F4.2 and Y37F4.4 appear to share a high scoring X box motif 51bp and 70bp upstream of the translation start site, respectively. Unfortunately, both are hypothetical genes with no functional annotation. Y37F4.4 consists of a single intron encoding a 96 amino acid gene product. A BLAST of both gene products against the non-redundant database at GenBank (Benson *et al.* 2005) results in no strong orthologues or protein domains.

Two of the genes identified with GFP::promoter expression data are expressed solely in ciliated neurons. The higher scoring of the pair is the orthologue of the recently identified human *BBS5* gene, R01H10.6 (Li *et al.* 2004). BBS5 was identified through creation of a flagellar apparatus basal body (FABB) proteome. The FABB proteome consists of proteins that are common to all organisms with flagella and absent in organisms without. Only two genes from the FABB proteome mapped to the *BBS5* region, one being the human orthologue of R01H10.6. The promoter::GFP fusion protein is expressed specifically in the ciliated neurons as expected (Figure 3A).

Table 5 High scoring X box hits.

Gene		X box Hit			*C. briggase* Orthologue		*Ce*GEP		Discovery In Previous Studies			
Transcript	Locus	Score	Dist.	Orien.	Score	Dist.	SAGE	GFP Fusions	A	B	C	D
F58A4.14		21.70	-50	+			N					
F56H9.4	*gpa-9*	20.45	-223	+			N				×	
R05H10.5		20.22	-835	-	21.92	-158	C					
Y45F10D.15		20.22	-143	+	13.18	-125	N					
F14D7.8		19.81	-163	+	nf		N					
ZK682.7		19.80	-305	-	nf		N					
ZC132.3		19.72	-70	-								
Y43F8C.12	*mrp-7*	19.71	-78	+			N	C				
Y43F8C.4		19.54	-369	+	nf		OC					
T05A7.10	*fut-5*	19.48	-71	+					✓			
F56A3.4	*spd-5*	19.30	-166	-	16.92	-57	C	N				
C15C8.1		19.28	-110	+	13.18	-112	OC					
C27H5.7b	*ict-1*	19.06	-205	+	15.06	-82	OC			✓		✓
T05C12.8		19.06	-786	-	21.70	-88	NC					
F46F2.5		18.88	-57	+	nf		NC					
Y37F4.2		18.85	-51	-								
Y37F4.4		18.85	-70	+								
K05F1.5		18.65	-223	+	13.92	-213	C					
T12B3.1		18.65	-172	-	18.65	-183	N	N				
R01H10.6	*bbs-5*	18.56	-51	+	20.78	-55	OC	CS		✓	✓	✓
M04C9.5		18.47	-966	+	13.27	-554	N	C				
C15F1.7b	*sod-1*	18.33	-633	+	nf		N					
F49E12.8		18.33	-209	+	16.23	-238						
C53D6.5		18.15	-674	+								
C27F2.1		18.04	-71	+	17.18	-62	C					✓
C09E8.1b		17.93	-396	-	13.19	-404						
F29A7.7		17.80	-78	+	nf		N		✓			
C06A12.4	*gcy-27*	17.70	-62	-				N				
F13H8.3		17.70	-197	-	17.70	-206						
Y37E3.5		17.64	-102	+	nf			CS				
C48B6.8		17.61	-67	+	16.91	-71	C	N				✓
F26A3.6		17.58	-797	-	13.59	-946	C	N	✓			
T02G5.3		17.44	-96	+	18.04	-94	C					
C05G5.1		17.40	-328	-	nf		N					
C25D7.3	*sdc-3*	17.40	-975	-	nf		C					
C26B2.4		17.40	-56	+	18.33	-112	N	N				
C43C3.3	*dyf-8*	17.28	-93	+	15.39	-62	C	C				
T26C5.5		17.27	-569	-	nf		N					
C23H5.3		17.23	-69	+	13.14	-140	C				✓	
M163.5		17.23	-534	-			N					
F57C7.4		17.18	-797	-	nf							
ZC132.9		17.18	-740	-	14.85	-801						

The upstream region of their *C. briggse* orthologues were scanned: An 'nf' means although an orthologue was found, no X box site was discovered; a blank space indicates that no orthologue was discovered. For SAGE data: blank = no SAGE data, an N = no SAGE expression, NC = no ciliated neuron SAGE expression, C = ciliated neuron SAGE expression, and OC = overexpression in cilated neurons. For promoter::GFP fusion data: blank = no expression data; N = not expressed, NC = no ciliated tissue expression, C = ciliated tissue expression, CS = ciliated tissue specific expression (based on reported expression, not photographs). Previous Studies: A ((Sandelin and Wasserman 2004)), B (Avidor-Reiss *et al.* 2004), C (Efimenko *et al.* 2005), D (Blacque *et al.* 2005)

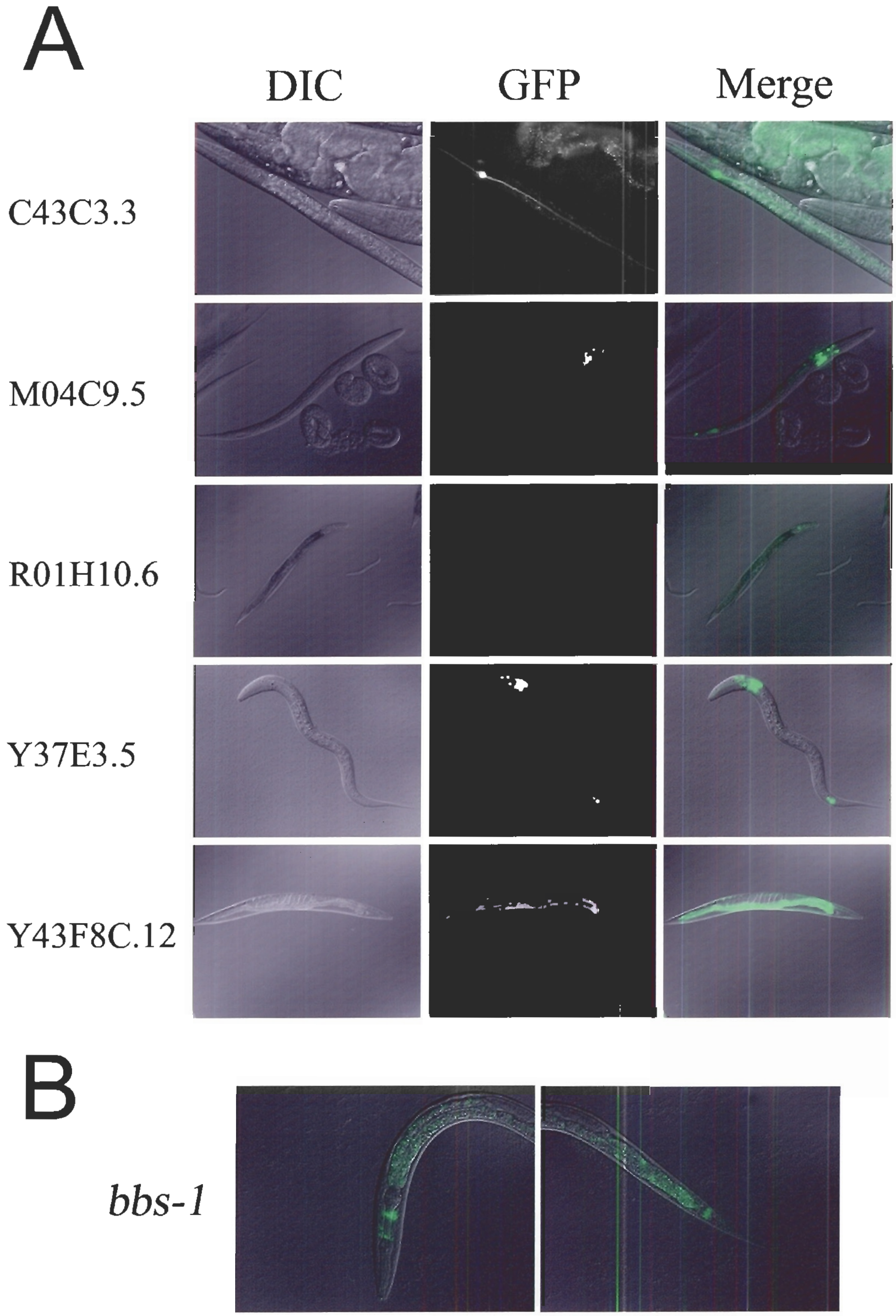


Figure 5 **The promoter::GFP expression pattern of candidate X box regulated genes**

(A) All genes identified in Table 5 with and promoter::GFP expression pattern.

(B) A typical expression pattern for a DAF-19/X box regulated gene, in this case *bbs-1* (Y105E8A.5).

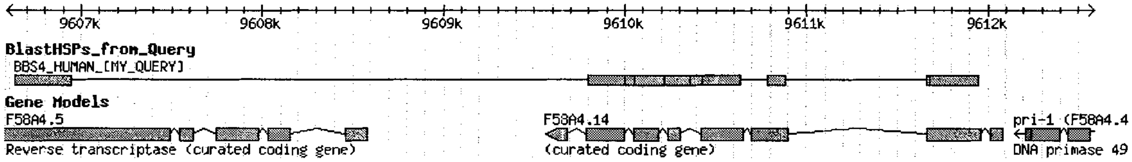


Figure 6 **BBS4 tBLASTn Against the *C. elegans* Genome Results**
from Wombase Server (Chen *et al.* 2005)

The second gene, Y37E3.5, a member of the ARL family of proteins, was a candidate *BBS3* orthologue in a recent study (Fan *et al.* 2004). ARL proteins are involved in biomembrane trafficking. Human othologues of Y37E3.5, *arl-6*, and *che-13* fell into the critical interval of *BBS3*. Sequencing of these genes in four independent BBS3 affected families resulted in the identification of *ARL6* as *BBS3*. The exact function of Y37E3.5 still remains unknown. However, despite this there is strong expression evidence that this gene is regulated by Daf-19. First, it has a high scoring X box motif (17.64) at close to 100 bp upstream of the translation start site. Second, the promoter::GFP fusion is reported to localize to tissues that are composed of or contain ciliated neurons (McKay *et al.* 2003). Closer examination of the image show that there is indeed amphid and phasmid expression (Collet *et al.* 1998). However, due to the strong expression in the image, confident exclusion of all non ciliated neurons in the nerve ring area is impossible.

Another interesting gene, M04C9.5, has an expression pattern is indistinguishable from Y37E3.5, yet has a strong X box hit over 9000 bp from the translational start site, much further than any proven X box site. The expression pattern for this gene is reported to include ciliated neuron containing tissue as well as the renal gland and the intestine. It would seem as though this visual similarity in the expression pattern of M04C9.5 and Y37E3.5 has no basis in coregulation.

Two additional genes are expressed in tissues containing ciliated neurons along with other tissues. However, the images show them unlikely to be regulated by Daf-19. C43C3.3, shows strong expression in the excretory cell and Y43F8C.12 appears to be strongly expressed in the intestine (Figure 3A). The raw data from this procedure is available in Appendix D.

Standardized Test Set

Recently, Tompa *et al* (2005) (Tompa *et al.* 2005) created a standard test set based on the TRANSFAC database and conducted an in-depth analysis of the publicly available computational tools for transcription binding site prediction. The intent of this analysis was to measure the current state of TFBS prediction as well as create a benchmark set of test sequences for future developed tools. In total, 13 motif discovery tools, of various methodologies, were analysed for their predictive power and accuracy. The dataset was carefully constructed based on the current understanding of TFBSs, which the authors admit, is limited. The described test procedure differs from the previous analysis in two major ways: 1) the analysis only takes into account the 'best' motif prediction, and 2) the test statistics are based on the nucleotide level of prediction, not the motif level. The tests do not allow any secondary information, such as comparative sequence analysis. OrBS was applied to all 52 available datasets (http://bio.cs.washington.edu/assessment/). Although allowed, no species specific background models were provided, allowing only higher-order background models where nucleotide count permitted (as previously described). This test set gives and unbiased analysis of the predictive power of OrBS.

Of the 156 datasets provided by the Assessment of Computational Motif Discovery Tools (ACMDT) website, OrBS predicted motifs in 114. The Assessment Score report provided by ACMDT is provided in Appendix C. Six of the datasets were rejected by OrBS for containing only one sequence. The two main statistics used in this assessment are the nucleotide level correlation coefficient (nCC) and the site level average site performance (sASP) (Tompa *et al.* 2005) of which OrBS scored a 0.177* and a 0.059 respectively. The data set is composed of sequences from four species: fly, human, mouse and yeast. Unlike the other tools which all performed best on the yeast datasets, OrBS preformed best on the mouse datasets with a nCC of 0.046, out performing eight of the fourteen tools tested in this category. No correct motifs were predicted for the fly datasets.

C. elegans Datasets

OrBS was not successful in prediction the PHA-4 binding site, TRTTKRY, in the set of genes upregulated in the pharynx. Examination of the motif chain states over the period of the programs execution show that while some chains did take on a state with a consensus sequence similar to the binding site it was typically for only a few sampling iterations. The first motif identified in all four replicate runs was a pyrimidine rich motif 9-12 bps in length, an average normalized OrBS score of 1.11. OrBS was also not successful in the identification of the UNC-86 binding site.

* The overall nCC score reported appears to be incorrectly calculated.

DISCUSSION

Algorithm

Background Model

In the original Gibbs sampler, both the background probability model and the motif model were evolving data structures. Lawrence *et al.* (1993) describe the background model as all the residues of the given sequences not contained in the current motif alignment. Therefore, whenever the alignment model changed, the background model also changed to reflect the changing motif. It is important to exclude the motif residues from the background when dealing with a small total number of total residues, because of the large effect on background probability model that the incorporation of these residues into the background model may have. Although this approach is statistically correct, it is not necessary when using large data sets. Where the total number of residues defining the motif model is minute compared to the number of residues defining the background model, as is the case with most gene promoter data, the effects on the probabilistic model of the removal of the motif from the background tends to be insignificant. In addition, the using only the sequence data provided to construct the background model is incorrect in both a statistical and biological sense. The sample size is usually much too small to construct a correct and robust model of background sequence. There is no evidence to lead one to assume that the background upstream promoter sequence of a set of coregulated genes is any different than the background upstream promoter sequence of the average gene. In fact, using only the promoters of interest may lead to an overrepresentation of certain sequences in the background model, leading to some motifs not being discovered.

For these reasons, the background probabilistic model used in OrBS can be generated from the provided sequence data or, preferably, generated from an independent data set of all upstream promoter regions for a particular species genome. In the case of the latter, the background model is provided to OrBS in a specified file format.

Scoring Function

Recently, much interest has been placed on how to incorporate additional information into *in silico* methods of regulatory elements prediction (Liu *et al.* 2004). Arguably, the most underused information is that of cross-species comparison. Currently, there are 15 completed eukaryotic genomes available at Ensembl(Hubbard *et al.* 2005), and there exist even more genomes partially sequenced. Like the genes they control, regulatory elements are under selective pressure to maintain a strong binding affinity with their associated transcription factor.

CompareProspector utilizes a WPID (window percent identity values) in an attempt to incorporate this information (Liu *et al.* 2004). Initially, CompareProspector only samples from those sites which have a WPID above a given threshold. As the algorithm runs, this threshold is gradually decreased to a lower value. I see a few problems with this methodology. First, the WPID threshold never reaches zero and therefore incorrect cross-species alignments or loss of a TFBS from the comparison species could result in the missing of sites with high resemblance to the current motif model. Second, the size of the WPID is irrespective of the current motif size and may reflect sequence conservation not included in the sampled site.

In OrBS I decided incorporate the bias towards conserved sequence into the core of the Gibbs sampling algorithm, the scoring function. This allows for the entire sequence space to be examined as well as only incorporating the conservation information of those sites in the motif model.

Another benefit of the scoring function modification approach is the ease of adding additional information. The activity of many TFBSs is positionally dependent. For example, the position of the known X box sites used in this research all exist between 60 and 160 bps upstream from the translation start site. I intend to include the option to positionally bias motif prediction in a later version of OrBS.

Motif Occurrence

There are a handful of existing methodologies to solve the motif occurrence problem. This includes both deciding when a sequence does not contain a motif as well as when a sequence contains multiple instances of a motif. The simplest is to ignore the problem and assume that all the input sequences contain a common motif. This is how the original Gibbs Sampler was implemented, but the authors were quick to suggest solutions to this problem (Lawrence *et al.* 1993). The solution this group came up with was outlined in the following instance of the algorithm, the site sampler (Neuwald *et al.* 1995). The column sampling algorithm treats the set of provided sequences as a single sequence and instead of sampling a specified number of sites from each sequence. The algorithm samples every site individually against the "null" model, the absence of a motif. While this method has been shown to be effective with protein domains, this methodology unnaturally biases greater motif occurrence in longer sequences. Initial tests with this method frequently predicted motifs with more than half the observed occurrences occurring in a single sequence. It is a natural assumption that there exists some uniformity in motif occurrence between co-regulated genes.

A second solution is to set two thresholds (Liu *et al.* 2001). A lower threshold (T_L), which is increased incrementally, sets a minimum score required for a site to be included in the sampling distribution. The second, higher threshold (T_H) sets a guaranteed motif alignment inclusion minimum. If a sequence has no sites with a score greater than T_L, the sequence is considered not to contain that motif. If the sequence has more than one site that scores greater than T_H, all those positions are added to the motif model. While Liu *et al.* (2001) claim using T_L decreases the convergence time of the algorithm, it actually increases the chances of converging, not on the optimal solution, but on a local maximum. Eliminating regions of the sequence set from analysis diminishes a Gibbs sampling algorithm to explore the entire dataset for the optimal solution by creating 'gaps' between maxima. For this reason, and the somewhat arbitrary nature of selecting a T_H, this methodology was not used in OrBS.

The final methodology assessed, and the algorithm implemented in OrBS, was the expected occurrence calculations outlined by Thijs *et al* (2001). This methodology combines the sequence length independence of the threshold methods with the sound statistics of the column sampler. Unfortunately, while the statistical basis of the algorithm is well explained, the implementation of these formulas into efficient code is almost unmentioned. The authors state that they used a modified Forward Algorithm, an algorithm typically reserved for calculating the probability a sequence was derived from a particular Markov Chain. It is unknown how the original authors implemented this methodology and the algorithm may vary dramatically from that described in the Methods and Materials section. However, as the authors claimed, the algorithm was implemented in linear time with respect to sequence length.

Motif Detection

Most previously described Gibbs sampler based motif prediction programs state that the algorithm will iterate over a motif until a maximum number of iterations or convergence has occurred (Aerts *et al.* 2003a; Liu *et al.* 2004; Neuwald *et al.* 1995; Wang and Stormo 2003). However, rarely do they define convergence or how the occurrence is observed. A major drawback of MCMC methods is the often difficult task of deciding when convergence has occurred and thus chain termination is possible. At convergence, a Gibbs sampler returns a sample from a distribution. The difficulties arise from the nature of the Markov algorithm that the sample will be generally correlated (Cowles and Carlin 1996). The two most popular convergence diagnostics in the statistical community are that of Gelman and Rubin (Gelman and Rubin 1992) and of Raftery and Lewis (Raftery and Lewis 1992). However, both these, and many other convergence diagnostics are mathematically complex and often require user input to decide when convergence has occurred. Most rely on the analysis of plots along with diagnostics reporter values in order to determine the time of convergence. In addition, there is much debate regarding the validity of each of these tests to the extent that it is most statisticians option that "a weak diagnostic is better than no diagnostic" (Cowles and Carlin 1996).

On the other extreme is the identification of a score maximum over a determined number of iterations. This seems to be the method of choice for most motif prediction algorithms. There is no confidence that the motif identified is the optimal solution and not a local maximum dependent on the starting alignment. This, however, is not as dire a situation as it seems. In the case of TFBS prediction, the algorithms are not designed to identify only the optimal solution but the set of maximal solutions that can be assigned as a common pattern with some level of confidence. The major problem with this approach is the decision of a minimum number of required iteration steps to reach convergence. Instead, a new, perhaps more intuitive approach was implemented in OrBS. In essence, the algorithm decides that "convergence" has occurred when a predetermined number of parallel running motif models converge to similar patterns.

The algorithm initializes multiple, parallel Markov chains with random starting states and, after an initial "burn in" phase, attempts to access convergence between the chains through a target matrix similarity test. Such tests are common in the literature for intra-database comparison of TFBS repositories (Hughes *et al.* 2000; Sandelin and Wasserman 2004; Schones *et al.* 2005; Wang and Stormo 2003). Three tests were analysed as candidate diagnostics: the Pearson correlation coefficient (Eisen *et al.* 1998; Schones *et al.* 2005), the Kullback-Leiber information/distance (Aerts *et al.* 2003b; Kullback and Leibler 1951), and the Pearson χ^2 test (Schones *et al.* 2005). Of these, the Pearson correlation coefficient was the measurement ultimately chosen. The Kullback-Leiber distance appears to be a natural choice because it is the same energy function used to create and refine the motif model. However, the lack of a constant normalization function to intuitively relate a given distance to a degree of similarity ruled out this approach. The range of both Pearson functions of negative one to one allows for the user to intuitively set a similarity threshold for convergence. Even though Schones *et al.* (2005) have shown the Pearson χ^2 test is a more robust similarity measure than the Pearson correlation coefficient, the benefits are overshadowed by the added complexity. In OrBS, the similarity measure is not concerned with distant family members or motifs with low information content and therefore the simpler, quicker Pearson correlation coefficient is sufficient. The simplicity of the calculation also allows OrBS to examine all alignments of the two matrices, overcoming any phase shift problems, with little effect to the programs running time.

The question of the minimum number of iterations required to identify the optimal solution as well as all members of the maximal set is another issue. Instead of implementing a minimum number of total iterations approach, OrBS uses a minimum number of iterations between convergence events approach. The reasoning is that while the program is still discovering motif patterns, there is no reason to stop the sampling procedure but continue until it is deemed that no further patterns exist. After a sufficiently large number of iterations pass without a convergence event, it can be assumed that no other convergence events will occur.

Dynamic Width and Phase Shift

The dynamic width and phase shift algorithm implemented in OrBS is an extension of the phase shift correction algorithm suggested in the original Gibbs sampler paper (Lawrence *et al.* 1993). It is somewhat puzzling that no other motif prediction algorithms based on this paper implemented this suggestion. In order to reduce any motif size bias a uniquely normalized information content was developed for this procedure. By examining the average KLI of the outermost nucleotide columns of all TFBSs stored in TRANSFAC, we attempt to distinguish between biologically informative nucleotide variance and random nucleotide variance.

Program Refinement

The current user interface for OrBS is designed to ease parameter refinement and thus simple and crude. The interface will eventually have to be updated to simplify use, either through reprogramming or the creation of a Perl wrapper and website. The latter is the favoured choice due to the ability to create multiple interfaces such as an inclusion of an alignment algorithm and orthologue database.

Convergence and Termination

In the first stage of refinement the default parameters for the number of parallel chains (-M), the alternate number of similar motifs required for a convergence event (-CN), and the Pearson correlation coefficient threshold (-CS) were chosen. The most obvious decision was the of the similarity threshold. A higher similarity is a better representation of actual chain convergence. The value of 0.75 was only tested in the case that a value 0.90 was too restrictive for the variation introduced into the motif by the random sampling procedure (Table 2). The results prove that this was not the case and in theory a threshold of 0.75 requires only 4 positions of 6bp motif to be designated as similar enough for convergence. When I examined the subset of 0.90 threshold parameters combinations that identify the X box, two candidate sets emerged. The combination of 5 parallel chains with a convergence minimum of 2 chains from this point on defined as parameter combination A and the combination of 10 parallel chains with a convergence minimum of 2 defined as parameter combination B. The combination of 10 parallel chains with a convergence minimum of 3 is perfectly selective for the X box motif. However, the X box is a highly conserved motif, and the incorporation of a few incorrect alignments has minimal effect on the motif model. In the case of a far less conserved motif, these parameters would most likely be too stringent. The added "incorrect" motif predictions in the chosen combinations of parameters are easily filtered out with a minimum score cut off. However, a minimum cut-off has will not be implemented until later in the development process in order not to miss any predictions.

Table 6 Maximum iteration between convergence events

	Maximum Iteration Between Convergence Events (-Ci)	**100**	**500**	**1000**	**2500**	**5000**	**7500**	**10000**
A	Motifs Predicted	0.67	1.00	1.00	1.33	1.67	2.67	3.33
	Average Raw Score of Motifs	19.12	18.88	18.88	16.21	15.15	13.19	12.71
	Average Number of Iterations Between Convergence	36	93	93	361	956	1804	2798
B	Motifs Predicted	1.00	2.33	2.33	4.67	7.00	9.00	14.33
	Average Raw Score of Motifs	15.50	13.65	13.65	11.61	10.97	10.67	9.89
	Average Number of Iterations Between Convergence	37	207	207	754	988	1126	1337

The run time of the algorithm is directly proportional to the number of parallel chains. In order to reduce running time, it may seem as though parameter set A would be a superior choice to parameter set B in an exhaustive search. However, examination of incremental maximum iterations between convergence events shows, in the case of the X box set, that set B surpasses, with a Ci value of 2,500 iterations, the number of motifs identified by set A with a Ci value of 10,000 (Table 6). In other words, although set B has twice as many parallel chains, it converges on motifs more frequently. For these reasons both sets A and B were chosen for the following noise/occurrence tests.

Occurrence and Noise

As seen in the results, Table 4, the occurrence algorithm was quite efficient at removing random, non-coregulated sequence from the sequence set. For both parameter sets A and B, the X box motif was detected consistently by OrBS for a datasets in which at least 50% of the genes contained a regulatory element. Although there is no literary reference to the noise removal efficiency for alternative motif prediction algorithms for comparison, this efficiency of OrBS in these tests seems adequate. It would be interesting to know what the typical co-regulation percentage of genes identified as co-expressed by various methods would is. Due to the natural distribution of the X box regulatory element, the performance of the occurrence algorithm was not observed in cases of multiple motif occurrences per input sequence.

Orthology

The orthology biased scoring function appears to successfully increase the prediction power of OrBS in the case of X box regulatory motif prediction. The X box motif was predicted consistently when only 33% of the input sequences contained the motif (Table 4). The resulting predictions were either comparable or better, based on the statistical tests used, than when no comparative information was provided. This further validates the use of comparative genomics in the discovery of novel transcription factors, even when the expression pattern of the putative orthologues is unknown. The phenotypic similarity of *C. elegans* and *C. briggsae*, despite the distant evolutionary separation, is a great aid to this inference. The sequencing of an additional three nematode species, *Caenorhabditis remanei, Caenorhabditis* n. sp. PB2801 and *Caenorhabditis japonica* is under way and will greatly benefit further study in the field of TFBS discovery. The inclusion of additional species will act to buffer any artefacts of the methodology used to generate the nucleotide orthology bonus of a given sequence as well as any natural divergence present between any two given species.

Two other recent attempts to incorporate comparative genomics information into a Gibbs sampling based regulatory element detection tool are CompareProspector (Liu *et al.* 2004), an expansion of BioProspector (Liu *et al.* 2001), and PhyloGibbs (Siddharthan *et al.* 2005). CompareProspector uses a naïve approach to incorporate this additional information. As previously discussed, only those motif positions which have a conservation score greater than a user defined threshold are considered valid motif positions. It would have been helpful if the authors compared the results of CompareProspector and BioProspector from a set of data in order to determine the benefit of the additional comparative genomics information.

PhyloGibbs, is not a conventional motif detecting Gibbs sampler. To create the phylogenetic information, the algorithm starts by aligning all ortholog groups using Dialign (Morgenstern 1999), to identify conserved blocks. Regions which are not conserved are treated as in a traditional Gibbs sampler. However, those regions that show conservation are treated quite uniquely. The algorithm makes the likely assumption that any putative motif that exists in a region of conservation must have evolved from an ancestral motif. The motif weighting function of the sampling step is expanded to incorporation this assumption. The authors state that this methodology accounts for non-functional conservation between closely related species, and the results show a marked increase in the efficiency of motif detection with the additional of phylogenetic information. Similar sophisticated comparative genomics information handling may increase the effectiveness of OrBS motif detection as well.

Performance on Known Sets

A training set of 14 genes known to contain the X box was used to aid in the development and calibration of OrBS. In almost all cases, OrBS performed exceptionally well on this set, typically predicting the width of the motif within two nucleotides of the actual size and up to a random sequence noise of 67%. The program was also successful in detecting the X box in all four runs of the 11 gene test set, indicating the program was not over-fitted to the training set in the refinement stages. These results validate the approach used to create this motif discovery tool. In addition to the program's predictive success on the test sets, the motif model derived from these test was efficient in the detection of additional X box motifs (Table 5). Most notably, the orthologue of the human BBS4 gene, F58A4.14, was identified as the top scoring gene. The fact that F58A4.14 was only recently annotated is quite perplexing. Only one of the papers describing the cloning of the first five BBS genes (*BBS1*, *2*, *4*, *6*, and *7*) attempt to identify orthologous genes in other species (Badano *et al.* 2003; Katsanis *et al.* 2000; Mykytyn *et al.* 2001; Mykytyn *et al.* 2002; Nishimura *et al.* 2001; Slavotinek *et al.* 2000). A literature search reveals that the first reference to the *bbs C. elegans* gene class, including *bbs-1, bbs-2, bbs-7, and bbs-8,* is the paper describing the cloning of *BBS8* (Ansley *et al.* 2003). However, there is no mention made to what process or evidence these orthologous relationships were determined. I can only assume that these assignments were made by automated annotation. This would explain why F58A4.14 remained undiscovered. To ensure that identification of this gene was possible at the time these papers were published, BBS4 was BLASTed against the Wormbase *C. elegans* genome, release WS100, the earliest release available. The results were identical to the previous attempt with WS140 (Figure 6). The only other cloned BBS gene not assigned a *C. elegans* orthologue is *BBS6/MKKS,* which is assumed to have evolved in the mammalian lineage (Li *et al.* 2004). Despite this, the *C. elegans* genome was scanned for orthologues of BBS6 in the same manner as with BBS4. As expected, no strong HSPs were observed (data not shown).

The only other gene with strong evidence that the gene is Daf-19/X box regulated is Y37E3.5. This gene had both a strong candidate X box site as well as an expression pattern characteristic of Daf-19 (Figure 5,Table 5). The gene contains no SAGE probe sights and thus no further expression data could be gathered. Y37E3.5 has no strong homology to any functionally annotated genes and therefore would be an interesting gene to study its role in ciliated cells.

The gene M04C9.5 is interesting because it shares an almost identical visual expression pattern to that of Y37E3.5 (Figure 5). However, the only resemblance of an X box in the upstream promoter region of M04C9.5 is over 900 bp from the translational start site, a distance far beyond any confirmed Daf-19 regulated genes. Also, it appears to be expressed in the renal gland cells of *C. elegans* (McKay *et al.* 2003). These two genes exemplify the idea that coexpression does not require coregulation.

The performance of OrBS on the known test sets was far from impressive. Most disappointing was the absence the PHA-4 binding site in the motifs predicted in the pharyngeal expressed genes, where other Gibbs samplers have been successful (Liu *et al.* 2004). Analysis of the sequences shows that 169 of the 199 pharyngeal expressed genes contained at least one match to the PHA-4 binding site consensus in the upstream region, well within the sequence noise levels in which OrBS is capable of detecting the X box motif. The sampling chains were observed to enter into the state with a similar consensus to the PHA-4 site, however, it was maintained only for a few iterations. This resulted in an absence of multiple chain convergence at such a state. The motifs that were identified tended towards repetitive, low complexity patterns, such as the poly-pyrimidine motifs reported in the results. A scan for this motif in the 5' upstream regions of the entire genome shows no overrepresentation in genes expressed in the pharynx over other tissues. It is likely the case that the abundant presence of these low complexity regions in the 5' sequence result in high overall similarity of the sequences. Many TFBS discovery algorithms mask repeats and low-complexity regions prior to analysis. Such an approach was not implemented in OrBS in order not to disregard any overlapping binding sites and the added orthology bonus was created to guard against such possibilities.

Another cause of poor performance is the dynamic width and phase shift algorithm. The relative difference in the information content score is small among the possible outcomes of the algorithm at any given step. The resulting change in the width and phase of the motif over several iterations therefore becomes almost random. In a method similar to genetic drift, this can cause the algorithm to "drift" away from a "good" pattern over time, especially with small or weak motifs. I propose that a possible solution would be not to compare the overall KLI score of the candidate outcome motifs, but rate the contribution of the surrounding sequence columns individually. An independent decision would be made for both the left and right side of the motif model to maintain the current alignment borders, or to expand or contract this border by a single nucleotide. Like the currently implemented algorithm, this would account for both dynamic width and phase shifts.

Tompa *et al.* (2005) provide a standardized and needed set of sequences and statistical tests for comparison of the presently available tools for TFBS discovery. The performance of OrBS in the standardised test sets was comparable to some of the currently available TFBS discovery tools. Most notably, the nucleotide performance coefficient (nPC) of OrBS was on par with Consensus (Hertz and Stormo 1999) and GLAM (Frith *et al.* 2004), which are both MCMC sampling based tools. The best performing algorithm was Weeder (Pavesi *et al.* 2004). This algorithm implements an oligonucleotide counting approach using a suffix tree to increase the speed of the exhaustive search.

Analysis of the Markov chains in both the pharynx and UNC-86 datasets reveals that although the correct motifs were not identified, all chains did at some iterations hold a consensus sequence related to the known TFBS. However, because no two chains were in this stage at the same iteration, no convergence event occurred and the motif was not reported. An alternative approach to the convergence methodology implemented in OrBS would be to let a single chain run for a given number of iterations, then select the "best" stage that the chain encountered as the top motif. To identify multiple motifs, the best motif is then masked and the procedure is re-executed. This "naïve" implementation was considered during the initial stages of the development of OrBS but the choice to go with the currently implemented approach was due to its better relation to MCMC statistics. In retrospect, this may have been the incorrect choice. If I examine only the UNC-86 data, the motif CAATGMAT was the highest scoring 8bp motif identified by OrBS. This is a close to the literary consensus binding sequence for UNC-86, AAATKCAT (Duggan *et al.* 1998), and almost identical to the motif identified by CompareProspector, CAATGCAT (Liu *et al.* 2004). The less than impressive performance of OrBS on these two data sets was expected. OrBS has so far has been designed to predict the occurrence of the X box motif with high fidelity. The inclusion of these additional TFBS test sets were only to determine the current state of OrBS as a general TFBS predictor.

CONCLUSION

The Orthology Biased Gibbs Sampling (OrBS) program is currently in a beta state. The program shows promise but also requires improvement. I effectively identified the X box in both the training and test sets as well as created a diagnostic motif capable of identifying additional X box regulated genes. Some of the modifications to the original algorithm, such as the inclusion of the orthology biased score, were successful at increasing the predictive power of the naïve Gibbs Sampler. However, others, such as the dynamic width algorithm and convergence methodology, were not. Despite these setbacks OrBS was able to identify the X box motif efficiently in high noise circumstances. After the suggested corrections are made, OrBS will be refined to identify additional known binding motifs. As previously stated, the methodology used in the creation of OrBS was to create a program that would perform extremely well on a single known TFBS. This is true for the X box regulatory motif. Now that this has been accomplished, I will sequentially refine the algorithm, using additional known TFBSs, to a more general predictor. Ultimately we wish to apply OrBS to co-expressed genes sets produced by the *Ce*GEP for the identification of novel *C. elegans* TFBSs.

APPENDICIES

Appendix A: OrBS Source Code

The source code for the OrBS program can be found on the accompanying CD in the "AppendixA" directory. Although designed to be platform independent, OrBS has only been compiled and tested on Linux systems using the GNU Compiler Collection (http://gcc.gnu.org/). The following is a table of the contents of the "AppendixA" directory.

Filename	Size (in bytes)	Short Description
dnaseq.cpp	4404	DNA sequence class implementation file
dnaseq.h	2529	DNA sequence class header file
gff_conv.cpp	10519	OrBS main program file
gibbs_conv.cpp	12901	OrBS parallel chain implementation file
gibbs_conv.h	5079	OrBS parallel chain header file
gmotif_conv.cpp	30851	OrBS motif implementation file
gmotif_conv.h	7306	OrBS motif header file
lrfloat.cpp	4757	Long range float implementation file
lrfloat.h	2091	Long range float header file
makeconv.sh	104	Bash script to compile OrBS executable
nfstream.cpp	3608	Fasta stream implementation file
nfstream.h	738	Fasta stream header file

All files are in text and can be viewed in any text editor / word processor program. Line returns have be converted to DOS/Windows format for ease of viewing.

Appendix B: Raw Data from X box Scan

The raw data used to construct Table 5 is available on the accompanying CD in the directory "AppendixB" under the filename "Xbox Scan Analysis.xls"(35,840 bytes). Unlike Table 5, the resulting hits in the promoter regions of the X box training set are included (blue text). The columns of the spreadsheet are as follows:

Column	**Description**
A	Gene Name.
B	Locus Name.
C	Sequence of X box hit.
D	Score of X box hit.
E	Distance from ATG of the X box hit.
F	Orientation of the X box hit.
G	Sequence of the *C. briggsae* X box hit.
H	Score of the *C. briggsae* X box hit.
I	Distance from ATG of the *C. briggsae* X box hit.
J	Sequence of the SAGE tag.
K	The SAGE tags position in the gene.
L	SAGE count in the FACS sorted ciliated neurons set.
M	SAGE count in the FACS sorted pan-neuronal cells set.
N	SAGE count in the FACS sorted muscle cells (replicate 2) set.
O	The number of other genes that share the same tag at the same position.
P	Identified in the Li *et al.* (2004) paper.
Q	Identified in the Avidor-Reiss *et al.* (2004) paper.
R	Identified in the Elfimenko *et al.* (2005) paper.
S	Identified in the Blacque *et al.* (2005) paper.
T	Strain number for the promoter::GFP fusion transgenic.
U	number of tissues with GFP expression that contain ciliated neurons.
V	number of neuronal tissues with GFP expression that do not contain ciliated neurons.
W	number of tissues with GFP expression in non-neuronal tissues.
X	Raw promoter::GFP Expression Data.

Appendix C: Gene Promoter Sets

Canonical X box Set: Y105E8A.5, F40F9.1a, F20D12.3, R31.3, F38G1.1, Y41G9A.1, F02D8.3, F59C6.7, T27B1.1, F33H1.1a, Y75B8A.12, T25F10.5, Y110A7A.20 and K08D12.2.

Noise (25%) Set A: B0207.5, C09G4.2b, E04F6.1, F08F1.1a, F23G4.t1, F41C6.1, and Y17G9B.7

Noise (25%) Set B: C17E7.5, C49C3.12, F53H2.1, T07D3.4, Y105C5B.9, Y53F4B.38, and ZK892.1b

Noise (25%) Set C: D1069.2, F01G10.3, H28O16.2, T04B8.5a, Y38E10A.5, Y46D2A.2, and Y73E7A.3

Noise (25%) Set D: B0511.9b, F09E5.11, K12C11.2, K12F2.2b, R74.4, T04B2.2, and T20F10.5

Noise (50%) Set A: B0564.3, C39D10.1, C40A11.1, C49A9.6, F13B12.3, F22H10.3, F54F11.1, K07A1.8, R09B5.4, W10C8.3, Y60A3A.8, Y71F9AM.6, Y73F8A.12, and ZK829.7

Noise (50%) Set B: C07C7.1, C16C2.1, C36C9.t3, D1054.7, F27C1.7b, K03A11.1, K05B2.5, K07C11.4, K08H10.2a, M04B2.7, T12E12.1, Y113G7A.4, Y76B12C.1, and ZK849.5

Noise (50%) Set C: C09B8.7c, F36H1.4b, H21P03.1, R13H4.6, R53.7a, T05A10.6, T06G6.3b, T20F5.6, T24F1.4, T27C4.4c, W05B5.1, W07A8.3, Y52B11A.11, and ZK1290.1

Noise (50%) Set D: C01G10.5, C07A12.4a, F07B10.3, F07C3.5, F09B12.6, F58B4.3, K01A11.3, K11G12.5, T04A8.14, T06C12.8, T07G12.5, T11B7.4d, Y69A2AR.11, and ZK550.2

Noise (67%) Set A: C01F1.1, C01F4.2a, C03A7.2, C03F11.2, C07A12.5a, C24D10.4, C31C9.1a, D2045.1, F01D4.5a, F21F8.1, F23A7.2, F26F12.4, F47B3.3, F57B10.8, T02C12.2, T04A8.14, T10H4.2, T20D4.13, T28F3.9, W05H7.3, W06G6.8, W08E12.8, Y40H4A.1b, Y50D7A.11, Y51A2D.10, Y54G2A.2a, Y62F5A.1b, and Y66H1A.3

Noise (67%) Set B: C15H11.7, C17G1.1, C17G1.4b, C23G10.2b, C33C12.1, C44F1.5, C53B4.1, DH11.5c, F01F1.1a, F09E10.6, F36A2.6, F41E6.4a, F41G3.6, F56A3.4, K05D4.8, R05D8.1, R06F6.8b, R09E10.3, R12E2.12, T12A7.4, T24E12.8, W04A4.5, W09D10.5, Y39A3A.5, Y50D7A.1, Y87G2A.12, ZC64.4, and ZK546.17

Noise (67%) Set C: B0353.1, C08F1.9, C17A2.1, F01G12.2a, F07A11.2b, F25B4.2, F25E5.9, F26F4.10a, F41B5.4, F53E2.1, H04M03.4, H06I04.t1, K08H10.9, M04D8.2, R13D7.9, T06E4.5, T08B2.4, T12B5.3, T20H4.3b, T26C11.4, T28H11.4, W02D3.8, Y41G9A.6, Y48G1BM.9, Y53F4B.26, Y67D8C.10b, Y71H2AM.9, ZC47.14

Noise (67%) Set D: B0410.1, C35B1.2c, C41D11.2, C48B6.6b, F18F11.1, F37B1.7, F57B10.8, H10E21.3a, H12I19.8, K08C9.6, M110.4a, M162.10, T06E8.1, T11B7.2, T19C4.9, T28D6.5a, W03A5.6, W07A12.8, Y49E10.5, Y49F6B.7, Y60A9A.1, Y62E10A.16, Y69A2AR.3, Y77E11A.t1, ZC455.5, ZK180.2, ZK418.2b, and ZK994.1

Appendix D: Results from Assessment of Computational Motif Discovery Tools

All data provided by the ACMDT as a result of the submission of the OrBS motif predictions are provided on the accompanying CD in the "AppendixD" directory.

Filename	Size (in bytes)	Short Description
ACMDT results.xls	203589	The results of the ACMDT analysis
ACMDT submission.txt	19662	Motif predictions submitted to the ACMDT

The column heading for the ACMDT results, as stated in Tompa *et al.* (2005), are as follows:

nTP	the number of nucleotide positions in known and predicted sites.
nFP	the number of nucleotide positions in known sites but not in predicted sites.
nFN	the number of nucleotide positions not in known sites but in predicted sites.
nTN	the number of nucleotide positions in neither known nor predicted sites.
sTP	the number of known sites overlapped by predicted sites.
sFP	the number of known sites not overlapped by predicted sites.
sFN	the number of predicted sites not overlapped by known sites.
nSn	nucleotide level sensitivity.
nPPV	nucleotide level positive predictive value.
nSp	nucleotide level specificity.
nPC	nucleotide level performance coefficient.
nCC	nucleotide level correlation coefficient.
sSn	site level sensitivity.
sPPV	site level positive predictive power.
sASP	site level average site performance.

BIBLIOGRAPHY

Aerts, S., G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. De Moor. 2003a. Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Res 31:1753-64.

Aerts, S., P. Van Loo, G. Thijs, Y. Moreau, and B. De Moor. 2003b. Computational detection of cis -regulatory modules. Bioinformatics 19 Suppl 2:II5-II14.

Ansley, S. J., J. L. Badano, O. E. Blacque, J. Hill, B. E. Hoskins, C. C. Leitch, J. C. Kim, A. J. Ross, E. R. Eichers, T. M. Teslovich, A. K. Mah, R. C. Johnsen, J. C. Cavender, R. A. Lewis, M. R. Leroux, P. L. Beales, and N. Katsanis. 2003. Basal body dysfunction is a likely cause of pleiotropic Bardet-Biedl syndrome. Nature 425:628-33.

Avidor-Reiss, T., A. M. Maer, E. Koundakjian, A. Polyanovsky, T. Keil, S. Subramaniam, and C. S. Zuker. 2004. Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. Cell 117:527-39.

Badano, J. L., S. J. Ansley, C. C. Leitch, R. A. Lewis, J. R. Lupski, and N. Katsanis. 2003. Identification of a novel Bardet-Biedl syndrome protein, BBS7, that shares structural features with BBS1 and BBS2. Am J Hum Genet 72:650-8.

Bailey, T. L., and C. Elkan. 1995a. Unsupervised Learning of Multiple Motifs in Biopolymers Uning Expectation Maximization. Machine Learning Journal 21:51-83.

Bailey, T. L., and C. Elkan. 1995b. The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol 3:21-9.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2005. GenBank. Nucleic Acids Res 33:D34-8.

Blacque, O. E., E. A. Perens, K. A. Boroevich, P. N. Inglis, C. Li, A. Warner, J. Khattra, R. A. Holt, G. Ou, A. K. Mah, S. J. McKay, P. Huang, P. Swoboda, S. J. Jones, M. A. Marra, D. L. Baillie, D. G. Moerman, S. Shaham, and M. R. Leroux. 2005. Functional genomics of the cilium, a sensory organelle. Curr Biol 15:935-41.

Blumenthal, T., M. Squire, S. Kirtland, J. Cane, M. Donegan, J. Spieth, and W. Sharrock. 1984. Cloning of a yolk protein gene family from Caenorhabditis elegans. J Mol Biol 174:1-18.

Brenner, S. 1974. The genetics of Caenorhabditis elegans. Genetics 77:71-94.

Brudno, M., C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow, and S. Batzoglou. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721-31.

Chen, N., T. W. Harris, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, P. Canaran, J. Chan, C. K. Chen, W. J. Chen, F. Cunningham, P. Davis, E. Kenny, R. Kishore, D. Lawson, R. Lee, H. M. Muller, C. Nakamura, S. Pai, P. Ozersky, A. Petcherski, A. Rogers, A. Sabo, E. M. Schwarz, K. Van Auken, Q. Wang, R. Durbin, J. Spieth, P. W. Sternberg, and L. D. Stein. 2005. WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. Nucleic Acids Res 33:D383-9.

Collet, J., C. A. Spike, E. A. Lundquist, J. E. Shaw, and R. K. Herman. 1998. Analysis of osm-6, a gene that affects sensory cilium structure and sensory neuron function in Caenorhabditis elegans. Genetics 148:187-200.

Consortium, C. e. S. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282:2012-8.

Cowles, M. K., and B. P. Carlin. 1996. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. Journal of the American Statistical Association 91:883-904.

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner. 2004. WebLogo: a sequence logo generator. Genome Res 14:1188-90.

Dawe, A. L., K. A. Caldwell, P. M. Harris, N. R. Morris, and G. A. Caldwell. 2001. Evolutionarily conserved nuclear migration genes required for early embryonic development in Caenorhabditis elegans. Dev Genes Evol 211:434-41.

Dougherty, E. C., and H. G. Calhoun. 1949. Possible significance of free-living nematodes in genetic research. Nature 161:29.

Duggan, A., C. Ma, and M. Chalfie. 1998. Regulation of touch receptor differentiation by the Caenorhabditis elegans mec-3 and unc-86 genes. Development 125:4107-19.

Dwyer, N. D., E. R. Troemel, P. Sengupta, and C. I. Bargmann. 1998. Odorant receptor localization to olfactory cilia is mediated by ODR-4, a novel membrane-associated protein. Cell 93:455-66.

Efimenko, E., K. Bubb, H. Y. Mak, T. Holzman, M. R. Leroux, G. Ruvkun, J. H. Thomas, and P. Swoboda. 2005. Analysis of xbx genes in C. elegans. Development 132:1923-34.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:14863-8.

Ewens, W. J., and G. Grant. 2001. Statistical methods in bioinformatics: an introduction. Springer, New York.

Fan, Y., M. A. Esmail, S. J. Ansley, O. E. Blacque, K. Boroevich, A. J. Ross, S. J. Moore, J. L. Badano, H. May-Simera, D. S. Compton, J. S. Green, R. A. Lewis, M. M. van Haelst, P. S. Parfrey, D. L. Baillie, P. L. Beales, N. Katsanis, W. S. Davidson, and M. R. Leroux. 2004. Mutations in a member of the Ras superfamily of small GTP-binding proteins causes Bardet-Biedl syndrome. Nat Genet 36:989-93.

Fire, A. 1986. Integrative transformation of Caenorhabditis elegans. Embo J 5:2673-2680.

Friedman, P., E. Platzer, and J. Eby. 1977. Species differentiation in *C. briggsae* and *C. elegans*. J. Nematol. 9:197-203.

Frith, M. C., U. Hansen, J. L. Spouge, and Z. Weng. 2004. Finding functional sequence elements by multiple local alignment. Nucleic Acids Res 32:189-200.

Fujiwara, M., T. Ishihara, and I. Katsura. 1999. A novel WD40 protein, CHE-2, acts cell-autonomously in the formation of C. elegans sensory cilia. Development 126:4839-48.

Gaudet, J., and S. E. Mango. 2002. Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. Science 295:821-5.

Gelman, A., and D. B. Rubin. 1992. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science 7:457-472.

German, S., and D. German. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6:721-741.

Gupta, B. P., and P. W. Sternberg. 2003. The draft genome sequence of the nematode Caenorhabditis briggsae, a companion to C. elegans. Genome Biol 4:238.

Haycraft, C. J., J. C. Schafer, Q. Zhang, P. D. Taulman, and B. K. Yoder. 2003. Identification of CHE-13, a novel intraflagellar transport protein required for cilia formation. Exp Cell Res 284:251-63.

Haycraft, C. J., P. Swoboda, P. D. Taulman, J. H. Thomas, and B. K. Yoder. 2001. The C. elegans homolog of the murine cystic kidney disease gene Tg737 functions in a ciliogenic pathway and is disrupted in osm-5 mutant worms. Development 128:1493-505.

Hertz, G. Z., and G. D. Stormo. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15:563-77.

Hubbard, T., D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. M. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, and E. Birney. 2005. Ensembl 2005. Nucleic Acids Res 33:D447-53.

Hughes, J. D., P. W. Estep, S. Tavazoie, and G. M. Church. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol 296:1205-14.

Kal, A. J., A. J. van Zonneveld, V. Benes, M. van den Berg, M. G. Koerkamp, K. Albermann, N. Strack, J. M. Ruijter, A. Richter, B. Dujon, W. Ansorge, and H. F. Tabak. 1999. Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. Mol Biol Cell 10:1859-72.

Katsanis, N., P. L. Beales, M. O. Woods, R. A. Lewis, J. S. Green, P. S. Parfrey, S. J. Ansley, W. S. Davidson, and J. R. Lupski. 2000. Mutations in MKKS cause obesity, retinal dystrophy and renal malformations associated with Bardet-Biedl syndrome. Nat Genet 26:67-70.

Kelly, A., and J. Trowsdale. 1985. Complete nucleotide sequence of a functional HLA-DP beta gene and the region between the DP beta 1 and DP alpha 1 genes: comparison of the 5' ends of HLA class II genes. Nucleic Acids Res 13:1607-21.

Kullback, S., and R. A. Leibler. 1951. On Information and Sufficiency. The Annals of Mathematical Statistics 22:79-86.

Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262:208-14.

Lenhard, B., and W. W. Wasserman. 2002. TFBS: Computational framework for transcription factor binding site analysis. Bioinformatics 18:1135-6.

Li, J. B., J. M. Gerdes, C. J. Haycraft, Y. Fan, T. M. Teslovich, H. May-Simera, H. Li, O. E. Blacque, L. Li, C. C. Leitch, R. A. Lewis, J. S. Green, P. S. Parfrey, M. R. Leroux, W. S. Davidson, P. L. Beales, L. M. Guay-Woodford, B. K. Yoder, G. D. Stormo, N. Katsanis, and S. K. Dutcher. 2004. Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. Cell 117:541-52.

Liu, X., D. L. Brutlag, and J. S. Liu. 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput:127-38.

Liu, Y., X. S. Liu, L. Wei, R. B. Altman, and S. Batzoglou. 2004. Eukaryotic regulatory element conservation analysis and identification using comparative genomics. Genome Res 14:451-8.

MacMorris, M., S. Broverman, S. Greenspoon, K. Lea, C. Madej, T. Blumenthal, and J. Spieth. 1992. Regulation of vitellogenin gene expression in transgenic Caenorhabditis elegans: short sequences required for activation of the vit-2 promoter. Mol Cell Biol 12:1652-62.

Marchal, K., G. Thijs, S. De Keersmaecker, P. Monsieurs, B. De Moor, and J. Vanderleyden. 2003. Genome-specific higher-order background models to improve motif detection. Trends Microbiol 11:61-6.

Mathis, D. J., C. O. Benoist, V. E. Williams, 2nd, M. R. Kanter, and H. O. McDevitt. 1983. The murine E alpha immune response gene. Cell 32:745-54.

McKay, S. J., R. Johnsen, J. Khattra, J. Asano, D. L. Baillie, S. Chan, N. Dube, L. Fang, B. Goszczynski, E. Ha, E. Halfnight, R. Hollebakken, P. Huang, K. Hung, V. Jensen, S. J. Jones, H. Kai, D. Li, A. Mah, M. Marra, J. McGhee, R. Newbury, A. Pouzyrev, D. L. Riddle, E. Sonnhammer, H. Tian, D. Tu, J. R. Tyson, G. Vatcher, A. Warner, K. Wong, Z. Zhao, and D. G. Moerman. 2003. Gene expression profiling of cells, tissues, and developmental stages of the nematode C. elegans. Cold Spring Harb Symp Quant Biol 68:159-69.

McKnight, S. L., and R. Kingsbury. 1982. Transcriptional control signals of a eukaryotic protein-coding gene. Science 217:316-24.

Menkens, A. E., U. Schindler, and A. R. Cashmore. 1995. The G-box: a ubiquitous regulatory DNA element in plants bound by the GBF family of bZIP proteins. Trends Biochem Sci 20:506-10.

Morgenstern, B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15:211-8.

Mykytyn, K., T. Braun, R. Carmi, N. B. Haider, C. C. Searby, M. Shastri, G. Beck, A. F. Wright, A. Iannaccone, K. Elbedour, R. Riise, A. Baldi, A. Raas-Rothschild, S. W. Gorman, D. M. Duhl, S. G. Jacobson, T. Casavant, E. M. Stone, and V. C. Sheffield. 2001. Identification of the gene that, when mutated, causes the human obesity syndrome BBS4. Nat Genet 28:188-91.

Mykytyn, K., D. Y. Nishimura, C. C. Searby, M. Shastri, H. J. Yen, J. S. Beck, T. Braun, L. M. Streb, A. S. Cornier, G. F. Cox, A. B. Fulton, R. Carmi, G. Luleci, S. C. Chandrasekharappa, F. S. Collins, S. G. Jacobson, J. R. Heckenlively, R. G. Weleber, E. M. Stone, and V. C. Sheffield. 2002. Identification of the gene (BBS1) most commonly involved in Bardet-Biedl syndrome, a complex human obesity syndrome. Nat Genet 31:435-8.

Neuwald, A. F., J. S. Liu, and C. E. Lawrence. 1995. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci 4:1618-32.

Nishimura, D. Y., C. C. Searby, R. Carmi, K. Elbedour, L. Van Maldergem, A. B. Fulton, B. L. Lam, B. R. Powell, R. E. Swiderski, K. E. Bugge, N. B. Haider, A. E. Kwitek-Black, L. Ying, D. M. Duhl, S. W. Gorman, E. Heon, A. Iannaccone, D. Bonneau, L. G. Biesecker, S. G. Jacobson, E. M. Stone, and V. C. Sheffield. 2001. Positional cloning of a novel gene on chromosome 16q causing Bardet-Biedl syndrome (BBS2). Hum Mol Genet 10:865-74.

Pavesi, G., P. Mereghetti, G. Mauri, and G. Pesole. 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res 32:W199-203.

Qin, H., J. L. Rosenbaum, and M. M. Barr. 2001. An autosomal recessive polycystic kidney disease gene homolog is involved in intraflagellar transport in C. elegans ciliated sensory neurons. Curr Biol 11:457-61.

Raftery, A. E., and S. M. Lewis. 1992. Practical Markov Chain Monte Carlo: Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. Statistical Science 7:493-497.

Reith, W., and B. Mach. 2001. The bare lymphocyte syndrome and the regulation of MHC expression. Annu Rev Immunol 19:331-73.

Riddle, D. L. 1997. C. elegans II. Cold Spring Harbor Laboratory Press, Plainview, N.Y.

Salmon, K., S. P. Hung, K. Mekjian, P. Baldi, G. W. Hatfield, and R. P. Gunsalus. 2003. Global gene expression profiling in Escherichia coli K12. The effects of oxygen availability and FNR. J Biol Chem 278:29837-55.

Salmon, K. A., S. P. Hung, N. R. Steffen, R. Krupp, P. Baldi, G. W. Hatfield, and R. P. Gunsalus. 2005. Global gene expression profiling in Escherichia coli K12: effects of oxygen availability and ArcA. J Biol Chem 280:15084-96.

Sandelin, A., W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32 Database issue:D91-4.

Sandelin, A., and W. W. Wasserman. 2004. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. J Mol Biol 338:207-15.

Schafer, J. C., C. J. Haycraft, J. H. Thomas, B. K. Yoder, and P. Swoboda. 2003. XBX-1 encodes a dynein light intermediate chain required for retrograde intraflagellar transport and cilia assembly in Caenorhabditis elegans. Mol Biol Cell 14:2057-70.

Schneider, T. D., and R. M. Stephens. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18:6097-100.

Schones, D. E., P. Sumazin, and M. Q. Zhang. 2005. Similarity of position frequency matrices for transcription factor binding sites. Bioinformatics 21:307-13.

Siddharthan, R., E. D. Siggia, and E. J. van Nimwegen. 2005. PhyloGibbs: A Gibbs Sampling Motif Finder that Incorporates Phylogeny. PLoS Computational Biology preprint

Signor, D., K. P. Wedaman, J. T. Orozco, N. D. Dwyer, C. I. Bargmann, L. S. Rose, and J. M. Scholey. 1999. Role of a class DHC1b dynein in retrograde transport of IFT motors and IFT raft particles along cilia, but not dendrites, in chemosensory neurons of living Caenorhabditis elegans. J Cell Biol 147:519-30.

Slavotinek, A. M., E. M. Stone, K. Mykytyn, J. R. Heckenlively, J. S. Green, E. Heon, M. A. Musarella, P. S. Parfrey, V. C. Sheffield, and L. G. Biesecker. 2000. Mutations in MKKS cause Bardet-Biedl syndrome. Nat Genet 26:15-6.

Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. J Mol Biol 147:195-7.

Spieth, J., K. Denison, S. Kirtland, J. Cane, and T. Blumenthal. 1985. The C. elegans vitellogenin genes: short sequence repeats in the promoter regions and homology to the vertebrate genes. Nucleic Acids Res 13:5283-95.

Spieth, J., M. MacMorris, S. Broverman, S. Greenspoon, and T. Blumenthal. 1988. Regulated expression of a vitellogenin fusion gene in transgenic nematodes. Dev Biol 130:285-93.

Stein, L. D., Z. Bao, D. Blasiar, T. Blumenthal, M. R. Brent, N. Chen, A. Chinwalla, L. Clarke, C. Clee, A. Coghlan, A. Coulson, P. D'Eustachio, D. H. Fitch, L. A. Fulton, R. E. Fulton, S. Griffiths-Jones, T. W. Harris, L. W. Hillier, R. Kamath, P. E. Kuwabara, E. R. Mardis, M. A. Marra, T. L. Miner, P. Minx, J. C. Mullikin, R. W. Plumb, J. Rogers, J. E. Schein, M. Sohrmann, J. Spieth, J. E. Stajich, C. Wei, D. Willey, R. K. Wilson, R. Durbin, and R. H. Waterston. 2003. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. PLoS Biol 1:E45.

Sulston, J. E. 2003. Caenorhabditis elegans: the cell lineage and beyond (Nobel lecture). Chembiochem 4:688-96.

Sulston, J. E., and H. R. Horvitz. 1977. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev Biol 56:110-56.

Sulston, J. E., E. Schierenberg, J. G. White, and J. N. Thomson. 1983. The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev Biol 100:64-119.

Swoboda, P., H. T. Adler, and J. H. Thomas. 2000. The RFX-type transcription factor DAF-19 regulates sensory neuron cilium formation in C. elegans. Mol Cell 5:411-21.

Thijs, G., M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. 2001. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17:1113-22.

Thijs, G., K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. 2002. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol 9:447-64.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 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 22:4673-80.

Tompa, M., N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. 2005. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23:137-44.

Wang, T., and G. D. Stormo. 2003. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics 19:2369-80.

Wingender, E., X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, M. Pruss, F. Schacherer, S. Thiele, and S. Urbach. 2001. The TRANSFAC system on gene expression regulation. Nucleic Acids Res 29:281-3.

Xue, D., M. Finney, G. Ruvkun, and M. Chalfie. 1992. Regulation of the mec-3 gene by the C.elegans homeoproteins UNC-86 and MEC-3. Embo J 11:4969-79.

Xue, D., Y. Tu, and M. Chalfie. 1993. Cooperative interactions between the Caenorhabditis elegans homeoproteins UNC-86 and MEC-3. Science 261:1324-8.

Zucker-Aprison, E., and T. Blumenthal. 1989. Potential regulatory elements of nematode vitellogenin genes revealed by interspecies sequence comparison. J Mol Evol 28:487-96.