TRANSCRIPTIONAL REGULATION AND *CAENORHABDITIS ELEGANS IN SILICO*

by

MICHAEL THORNE

B.Sc., The University of British Columbia, 1993

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Genetics Graduate Program)

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

August 2001

~~Department of~~ GENETICS GRADUATE PROGRAM

The University of British Columbia
Vancouver, Canada

Date August 27th, 2001

Abstract

With both the complete sequences of multicellular organisms as well as the emerging results from genomic scale expression experiments at our disposal the incentive to construct models detailing the interactions of gene products is greater than ever. An integral understanding and input into any such model lies in the combinatorics of the transcription factors that result in gene expression. Footprinting experiments have led to the assumption that a specific transcription factor binds exclusively to a strongly conserved sequence of nucleotides. Consequently, the search for transcription factors has been simplified to the problem of determining the motif to which each transcription factor may bind.

Caenorhabditis elegans, with its essentially complete genome sequence and comprehensive annotation is currently the best *in silico* model to study transcriptional regulation in a multicellular organism. Gathering between 200-2000 bp of the upstream region of all genes unlikely to lie within a polycistronic transcript, a number of different approaches to finding candidate motifs have been applied. These include the study of over-represented oligonucleotides in the above mentioned dataset vis-a-vis the whole genome, the examination of signal distribution in the dataset, the comparative genomic approach of phylogenetic footprinting with *Caenorhabditis briggsae*, and analyses based on the results of gene expression technologies.

The cataloguing of motifs found through whatever means neccesitates a method of organization as well as a ranking according to their possible biological relevance. Grouping into a large matrix all the potential motifs on one axis and the genes they lie proximal to on the other eliminates positional and ordering information but enables one to draw on techniques from graph theory and mutivariate statistics, as well as providing the ability to cluster genes based on common transcriptional profiles. Such methods allow one to extract information about composite motifs and points to their potential use in determining, when looking at sets of coregulated genes, the underlying control mechanisms.

Contents

List of Tables

List of Figures

Introduction

Measure everything that is measurable, and make measurable everything that is not yet so.

- Galileo Galilei

René Thom, in his 1972 treatise "Stabilité Structurelle et Morphogénèse" [121], made the following comment on disciplines that have in the past proven uneasy bedfellows: "... conversely, if some disciplines, like social sciences and biology, resisted mathematical treatment for so long, even if they have succumbed, this is not so much because of the complexity of their raw material, as is often thought (all nature is complicated), but because qualitative and empirical deduction already gives them sufficient framework for experiment and prediction." By mathematical treatment, Thom is not referring to the use of quantitative assessment - this is a staple in the field of biology - but rather the use of mathematics to provide a description of biological phenomena, allowing for prediction and the analysis of behavioural properties. While it remains a daunting challenge to represent the morphology of aggregate cells when the aggregate constitute a multicellular organism, real insight has been made in modelling at the finer resolution of internal cellular processes and cell signalling.

The current excitement over gene expression technologies in determining the components involved in cellular processes testifies to the enormous strides in knowledge we are making. The value of predictive modelling, however, is not in hindsight description but in its ability to direct experiment. Or, in the boldest view, make experiment redundant. The current revolution in biology as a consequence of the sequencing of complete organisms has the potential to deepen this direction toward a predictive capacity. Following Zuckerkandl and Pauling's paper of 1962 on the relationship of globin genes from different organisms [142], the generation of vast amounts of sequence data has had the most immediate impact through inference by homology, and shall continue to do so. But it is the dynamic play of the gene products that has the largest role in the determination of an organism. Consequently it is the union of an organisms' complete DNA sequence with mRNA and protein expression technologies that will play the greater role in building predictive models.

Differential equations, the traditional method for depicting dynamic relationships, are always somewhat of a blanket to an underlying discrete combinatorics. Similarly, models exploiting gene expression data are in an equivalent relationship vis-a-vis the underlying combinatorics of the transcription factors of the regulatory systems. It is the vast orchestration of the transcription factors that, emerging from the studies of Jacob and Monod [51], provided hints of the intricate control mechanisms determining an organisms development. We have travelled a long way from the early work on the λ repressor [94], still the best understood regulatory system, and while many of the paradigms from bacterial systems naturally extend into the eukaryotic realm, the increase in complexity will likely be mirrored by an increase in the regulatory logic [26].

Much has been made of the results of experiments to determine the specific regions of

DNA to which transcription factors interact, the clearest of these being the footprinting method [17]. From the compilation of a number of DNA-protein interactions, an assumption has been promulgated that the determination of transcription factors, which play a role in gene regulation, may be reduced to whether the sequence to which the protein binds is in an effective position. This assumption has consequently taken the study of which transcription factors play a role in regulation into the computational arena, where motif searching has become one of the *de rigueur* "unsolved" problems. There are two distinct but intricately related problems motivating the community in this search for binding site motifs. The first is to find these motifs and to associate them with their affected gene, a process of cataloguing. The other is to use the knowledge of these motifs in lieu of the transcription factors as a basis for determining the logic underpinning the dynamics of gene expression. The second approach remains suspect in that there exists no proof that a consistent one-to-one association is feasible. The search for computational rules demonstrating the association of binding sites with specific transcription factors and further, those binding sites that are consequential for expression owing to their position, is bound to bring to mind the attempts made to find rules of binding based on the amino acid composition of motifs found among the families of transcription factors, a trickle of such papers still to be found in the journals [27] [77] [116]. One cannot but be reminded of those intrepid souls who sought and still seek to square the circle. The prognosis is not necessarily so dire however, in that a rules based approach may yet find a renaissance. It may simply be a question of timing, of a problem not yet being, as François Jacob would phrase it, within the envelope of the possible. While the problem addressed in this thesis is enticing, because easily stated and easily studied, the real dilemma lies in the verification.

In considering transcriptional regulation, the most fundamental step in a gene products' regulated life, I am drawn not to the reasons why an *in silico* deconstruction would work, but why it won't, or at least prove problematic. It is these reasons that shall, if truly of import, be the information that in the future must be incorporated into a method of detection, whether epigenetic or not. The list is hardly exhaustive: the effects of methylation, currently addressed by a public consortium [45]; conformational properties, a nicely documented case being the situation in which an *in vivo* footprinting result could not be replicated *in vitro* [17]; indications of heretofore unexpected mRNA coding complexity, such as with the *Drosophila melanogaster mod(mdg4)* [68]; enhancer blocking by insulators [11]. As I run through these problematic situations I think of current gene prediction programs that take into account proximal regulatory elements, lending greater weight to the decisions, and wonder how future prediction programs might incorporate all genomic knowledge into models analogous to Hidden Markov Models to allow for increased informative decisions. Would it be possible to predict whether a hexamer 50,000 base pairs away from a gene plays a role otherwise?

It is a striking testament to the fundamental change in biological investigation, along the lines of Jacques Monod's dictum that "what is true of *E. coli* is true of elephants," that I mention lastly, I shall not go so far as to say "merely in passing," the organism used as the model in this study. Why *Caenorhabditis elegans*? Simply because it is the most thoroughly sequenced multicellular organism. For proof of principle more headway might have been made with *Saccharomyces cerevisiae*, especially given the quantity and detail of expression data already compiled. But the potential payoff is greater with *C. elegans*. With *in silico* models there is no greater difficulty in kind between analyzing the genome of a single celled over a multicelled organism, the same techniques hold. And if there is an ability to derive pertinent facts concerning the elements underlying a complex system, then the more complex the system, the bigger the payoff. But *C. elegans* is an extraordinary model in its own right, well beyond the mere priority of its sequences' completion. Although taking the stage as a model organism for genetic research quite late, *C. elegans* has a number of firsts to its credit, the most profound of which has been the complete elucidation of its cell lineage [115]. Acquaintance with this free-living nematode worm as an object of scrutiny stretches back at least 100 years and even before its consideration as a genetic model some cytological facts, such as chromosome number, had been determined [85]. The stories surrounding Sydney Brenner's choice of *C. elegans* owing to its short life cycle, ease of handling, and transparent body have become legendary [100]. Far from expected from Sydney's early studies was its future role in elucidating the processes involved in apoptosis [46] and aging [63], the discovery of genetic tools such as RNAi [36], and its use as a testbed for the techniques of mapping and sequencing [134] [19] instrumental to the revolution we have the pleasure to be living through.

Chapter 1

Fits and Starts

Analyses typified by the inaugural papers on any new genome necessarily draw on the state of the art in any given area, such as gene prediction, and it is a testimony to the paucity of general principles in regulatory analysis that in the two year hiatus between the publication of the two genomes of *Caenorhabditis elegans* [19] and *Homo sapiens* [70], during which *Drosophila melanogaster* [2] was also completed, that very little ground has been gained even with the vast amount of expression data available. Indeed it has been stated that no significant progress in general principles has been made in the twenty years this problem has been investigated [118]. Current methods require the knowledge of an experimentally determined motif and, through a position weight matrix (PWM), one is able to say whether such a motif resides in a region proximal to a gene or not, certainly not whether the site is functionally relevant. In a companion paper to the publication of the *C. elegans* genome, discussing the predicted Zinc Finger genes, Clarke and Berg [21] considered significant both the distribution of a consensus TRA-1 binding site throughout the genome as well as the number of occurrences of the site upstream of the set of predicted genes. A similar analysis is incorporated in the *H. sapiens* publication [1]. While a few studies have proven successful in detailing the relationship between regulatory elements in eukaryotic systems, resulting in the regulatory module hypothesis [26] [139] [133], the problem from a genome perspective remains in its infancy.

The approaches used in this study are all extremely simple. As highlighted in the introduction the incorporation of increasingly more complicated scenarios is certainly a goal, but we are currently facing the first hurdles. The problem of searching for strings of negligible length in the immense search space of the genome need hardly be stated and therefore of vital importance is the ability to reduce the noise in order to increase the signal. In truth we cannot afford to reduce any search space. Transcription factors have an effect from introns, positions downstream of the 3´ untranslated region (UTR), exonic regions, and sites any number of base pairs upstream of the 5´ UTR [73]. Realistically, however, we must target the area containing the densest concentration of functional signals, immediately 5´ of the UTR [73]. The question of the average size of the UTR is also one not fully resolved in *C. elegans* and has necessitated utilizing the region 5´ of the ATG translational start site. While the occurrence of polycistronic transcripts in *C. elegans* [109] may provide a heretofore unexpected but valuable source of analysis (see Chapter 2), they also complicate the search for binding sites and are best excluded.

With the dataset generated as described in Appendix A.1 one can begin, so to speak, where Clarke and Berg left off [21]. The occurrence of a number of potential binding sites, both experimental (*in vivo* and *in vitro*) and computationally determined by base similarity with *Caenorhabditis briggsae*, has been determined, in an exact manner, in both the whole genome and the dataset, as well the distributions of occurrences in the upstream regions.

Binding Site	Associated Gene/TF	Reference	Total	Ratio	Max
TCCCYHRRGRRV	daf-7/UNC-3/OLF-1	D.Riddle (pers. comm.)	864	0.1632	4
RTAAAYA	FREAC-2,-3,-4,-7	M. Hellqvist-Greberg (pers. comm.)	105650	0.1476	39
TATAA	hsp-16 / basal	McGhee and Krause, 1997	306064	0.1472	32
CAGCTG	hlh-1 / MyoD	McGhee and Krause, 1997	39824	0.1246	12
AAATTCAT	mec-3 / UNC-86	McGhee and Krause, 1997	17668	0.1394	51
TGGGWGGTC	TRA-1L	McGhee and Krause, 1997	644	0.0963	3
RTCAT	SKN-1	McGhee and Krause, 1997	417174	0.1399	29
CTAAAAATA	MEF-2	McGhee and Krause, 1997	6512	0.1402	5
TAAAGTGGTTGTGTG	CEH-22	McGhee and Krause, 1997	2	0.5	1
RTGGGAA	LAG-1	McGhee and Krause, 1997	31372	0.1117	16
TGTCAAT	vit-2 / 6(VPE1)	McGhee and Krause, 1997	21780	0.1354	45
CTGATAA	vit-2 / 6(VPE2)	McGhee and Krause, 1997	19844	0.1530	8
TTTTCAGR	SL1 RNA	McGhee and Krause, 1997	64396	0.1255	19
GAGTATCNNNNNCTCTTC	her-1	Streit *et al.* 1999	6	0.1667	1
GAGTATCTAAGTCTCTTC	her-1	Streit *et al.* 1999	2	0.5	1
TTTGACCTT	mel-32	Vatcher, 1999	1184	0.1993	3
CAAACTACT	mel-32	Vatcher, 1999	992	0.1109	2
TCTTGTTTGCAACAA	mel-32	Vatcher, 1999	2	0.5	1
TGATCGATA	mel-32	Vatcher, 1999	720	0.1528	2
GCTTTTCTCTC	mel-32	Vatcher, 1999	132	0.1515	1
TTTTTTGTTTTT	mel-32	Vatcher, 1999	2972	0.1511	6
CAAGTGTCACC	his-24	Périer *et al*, 2000	34	0.1177	1
CATCAGATTCG	his-12	Périer *et al*, 2000	46	0.1739	1
TCTTCATCTCA	his-10	Périer *et al*, 2000	162	0.2469	1
ATAGAGTTCTC	msp-56	Périer *et al*, 2000	38	0.3158	1
CACTTGGCTTC	col-12 / F15H10.1	Périer *et al*, 2000	56	0.125	1
CACTTTATTTC	col-13 / F15H10.2	Périer *et al*, 2000	154	0.1753	1
ACGGTTCAGCC	vit-2	Périer *et al*, 2000	12	0.25	1
ACTCTCGCAAT	vit-5	Périer *et al*, 2000	22	0.2727	1
ACTCGGTCACT	vit-6	Périer *et al*, 2000	34	0.2647	1
CGAGCAGAAAG	cal-2	Périer *et al*, 2000	48	0.0625	1
GGCGGGTGTAT	kin-2	Périer *et al*, 2000	4	0.5	1
GGGTATCAATT	kin-2	Périer *et al*, 2000	28	0.1786	1
AAACAACATTC	hsp-16K-1	Périer *et al*, 2000	184	0.1630	1
AACCAATACAC	hsp-16K-48	Périer *et al*, 2000	62	0.1290	1
AGCTCAATTTG	mtl-1	Périer *et al*, 2000	134	0.0896	1
GAATCAAGCTT	mtl-2	Périer *et al*, 2000	60	0.1667	1
AATAACGTGTT	casein kinase II	Périer *et al*, 2000	134	0.1418	1
TRTTKRYTYS	pha-4	Gower *et al*, 2001	24332	0.1532	5
WGATAR	GATA-1	Gower *et al*, 2001	231084	0.1377	17
GGATTA	unc-25	Eastman *et al*, 1999	35468	0.1223	12
TAATCC	unc-25/unc-47	Eastman *et al*, 1999	35382	0.1463	14

Table 1.1: **Potential binding sites culled from the literature. The column labelled Total is the total number of exact occurrences of the binding site in the whole genome while the Ratio column is the proportion found solely in the generated dataset of upstream regions. The column labelled Max is the maximum number of sites in the representative upstream region. See Appendix B for interpreting the ambiguity codes in the binding sites.**

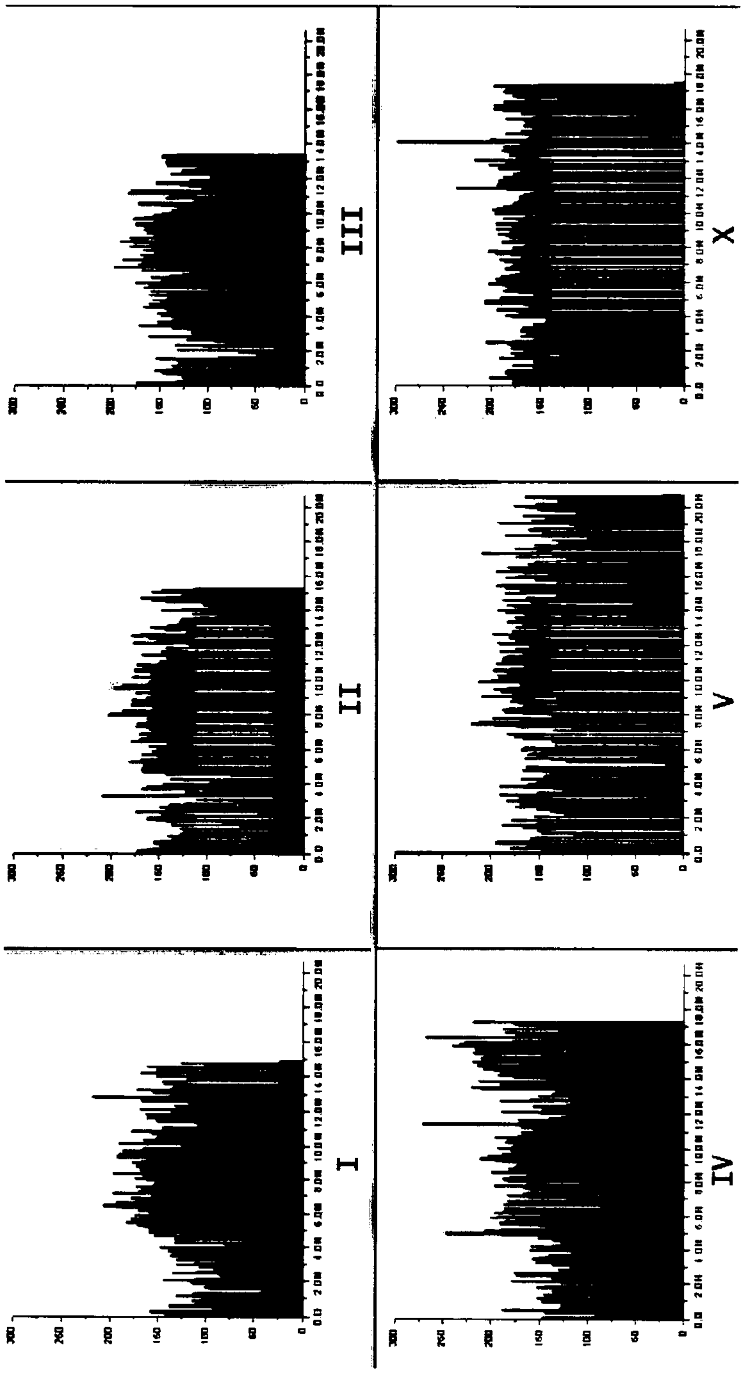


Figure 1.1: **The distribution of the TATA box throughout the *C. elegans* genome. Normalized along the horizontal axis by the size of chromosome V, each bin represents 2KB. The Y-axis is the number of occurrences of TATAA in the given bin.**

Shown in Table 1.1 is their occurrence in the whole genome, the ratio found in the dataset and the maximum number of occurrences found in the representative upstream region. It is worth while keeping in mind when looking at the table that the dataset represents 26% of the whole genome. Even so, little emerges from this picture but rather draws me to consider what is known about competition and specificity. Ptashne, in his book *A Genetic Switch* [94], in surveying the results from *E. coli* on specificity and the amount of free repressor deduces that 99% of the repressor is not free in solution, but bound to non-specific operators. He closes by considering the eukaryotic case no different in kind. Does it matter then whether we are dealing with non-specific bound factors or specific non-functional bound factors, when so many are inconsequential?

The TATA-box is, of course, quite central. It has been depicted in both its genome-wide distribution (Figure 1.1) and its distribution of occurrences in the upstream regions of the dataset (Figure 1.2(A)). From Figure 1.2(B), however, emerges a point of possible significance. Representing the positional distribution of TATAA sites over the set of upstream regions shows a peak $\approx 50bp$ from the ATG start site. This hints at two things. A possible indication of the average size of the $5'$ UTR, as well as a depiction of functional signal. Looking at this figure I am drawn to restate the usual rendition of motif searching from a signal to noise problem in the obvious sense to a problem where the noise is identical to the signal. In this restatement, position takes centre stage.

Since the availability of whole genomes, replete with gene predictions, it is a very natural question to ask, but possibly a dangerous one to assume, whether motifs that play a role in binding transcription factors are in greater abundance in the regions where they are of import. Figure 1.3(A) depicts the frequency of all hexamers in their ratio of occurrence in

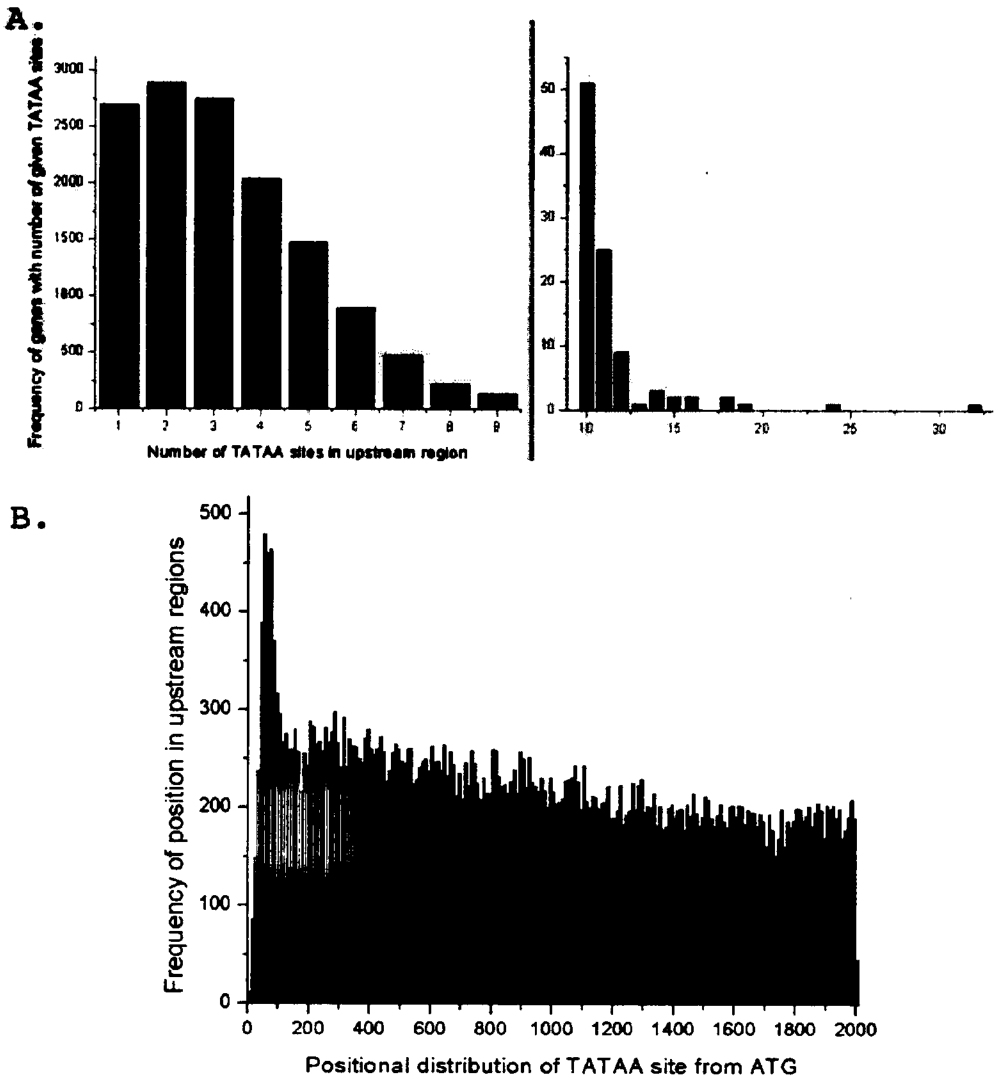


Figure 1.2: **Distribution of the frequency of the quantities of TATAA in the dataset of 15,525 upstream regions (A) and their positional distribution (B), determined by plotting the start site of each TATAA over all 15,525 upstream regions, where 0 depicts the ATG translational start site.**

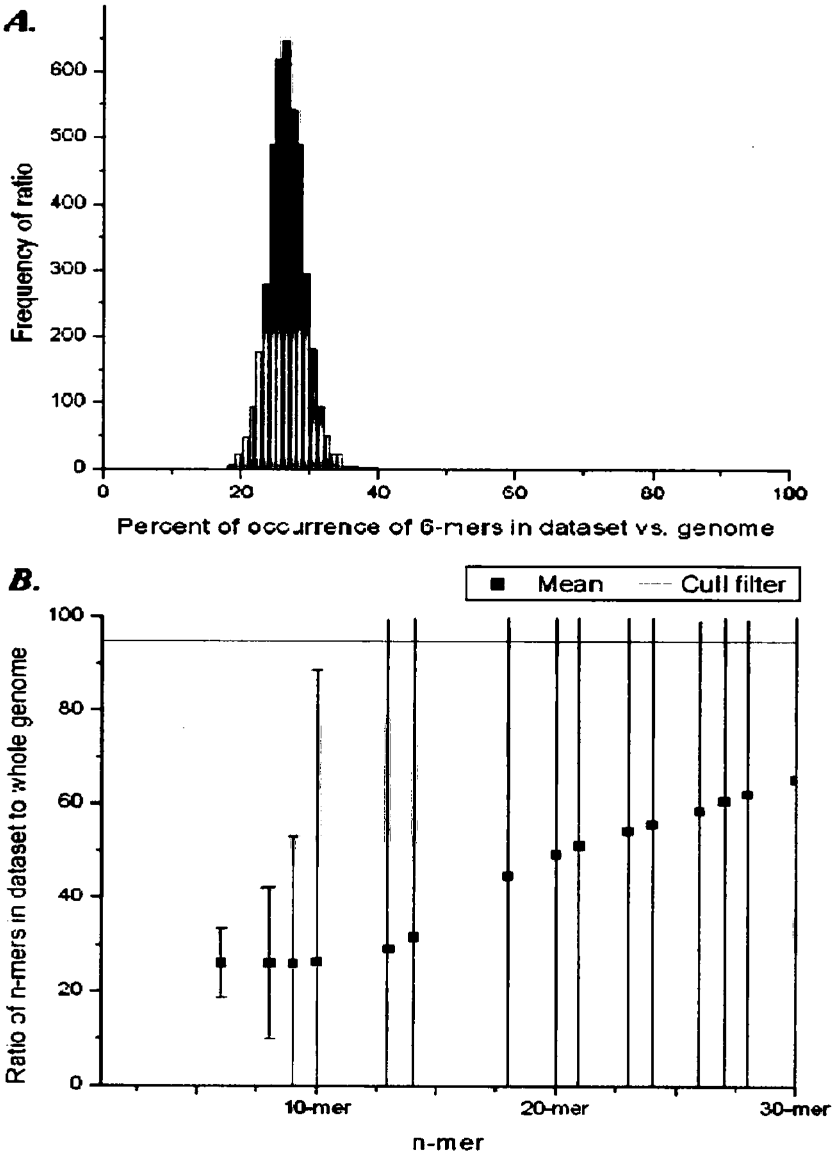


Figure 1.3: **The distribution of the ratio of hexamers (A) and from hexamers to 30-mers (B) in the upstream regions vs. the whole genome. Generated by determining exact quantities of all n-mers in both the dataset and the whole genome and plotting the frequency of the ratios. Those motifs above the line at the 95% mark in (B) were chosen to aid in building PWM's as shall be described in Chapter 3.**

the generated dataset to the whole genome. The mean is $\approx$ 26%, the same as the ratio of the dataset to the whole genome, a definitive statement that no hexamers involved in transcriptional regulation in *C. elegans* are more representative in the regions where they presumably play a role. Figure 1.3(B) pursues this line up to 30-mers. As we increase the size of the n-mer an interesting question takes root. While it is natural that large n-mers playing a role in transcription will have increased representation in the dataset - a consequence of every 13-mer in theory able to reside within *C. elegans* but not even half of all 14-mers - it is also true that as a motif increases in size the variability in nucleotide composition will also increase [94].

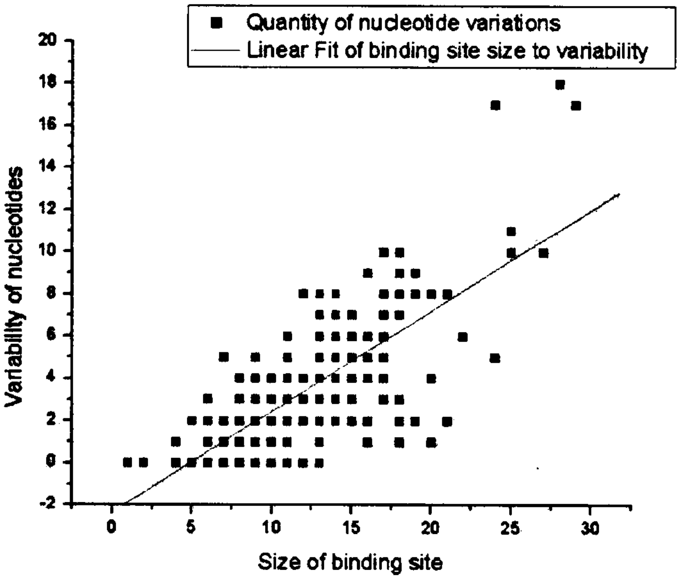


Figure 1.4: **Linear fit of variability of nucleotides vs. binding site size. Generated by stripping all N's, representing any of the possible four nucleotides, off the edges of the 280 motifs in TRANSFAC and plotting the size vs. the variability in the motif.**

This question of the variability of nucleotides is worth investigating, but reservations apply. Footprinting studies have determined that it is often only necessary to mutate one critical base pair to cripple a protein-DNA interaction whereas a number of other base pairs may be variable with impunity [3]. Regardless, asking this question through analysis of the transcription factor database, Transfac [137], by gathering all 280 motifs, stripping any N's, representative of any possible base pair, from the edges, and plotting the number of ambiguous bases against the size of the binding site yields Figure 1.4, an increase that is roughly linear.

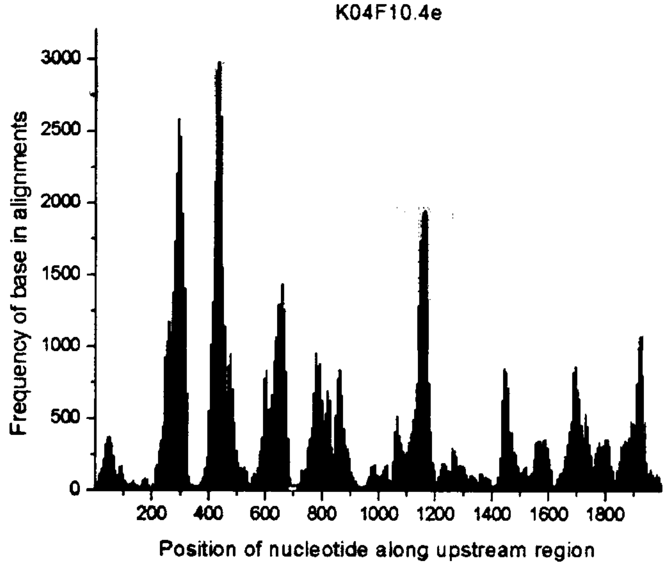


Figure 1.5: **Distribution of the base pairs of the upstream region of *bli-4* in alignment with all other upstream regions in the dataset. See Appendix A.2.**

Knowledge of the variability of nucleotides as the binding site grows is valuable when determining generalized searching strategies. One such approach has been motivated by the idea of illuminating the nature of the upstream regions comprising the dataset. While it was clear from Appendix A.1 that the dataset represents no great change in nucleotide composition, it is likely that a greater density of functional binding sites occur in this set.

Given that, what sort of properties would emerge by a vast cross-comparison of any given upstream region against all other upstream regions in the dataset, utilizing an alignment mechanism customized to preference the unique properties of binding sites? The alignment of binding sites, regardless of whether there are mismatches, tend toward having consistent spacing. Consequently, a modification of the Smith-Waterman algorithm, removing the scoring mechanism for vertical or horizontal movement, would retain only diagonal scores (see Appendix A.2). This query on the nature of an upstream region vis-a-vis all other upstream regions is an attempt to ask what is common, what is unique? Figure 1.5 shows the disparity in the alignments over the upstream region of the *bli-4* gene. While it might have been expected to be awash in noise, there is clearly strong demarcation. This procedure proved to be computationally taxing and was not followed through with enough upstream regions to gauge any possible emergent properties but it does point to regions of greater representation for the set as a whole. The immediate reaction is to consider these regions of greater representation as ubiquitous binding sites. This may well prove to be so but the regions examined, for instance in *bli-4*, showed no correlation to those sites known to have functional relevance.

1.1 Phylogenetic Footprinting

Widespread opinion holds that the initial tasks set out in this study will prove redundant in a short while owing to the technique of phylogenetic footprinting [93]. As a model, *C. elegans* is far more powerful thanks to our knowledge of another free-living nematode, *Caenorhabditis briggsae* [37] [43] [114]. The morphological proximity of the two species, despite the estimates of divergence between $\approx 23-40myr$ [119], is such that wonderful stories of researchers confusing the two species, or thinking them to be one and the same, were finally laid to rest by mating experiments, their inability to cross ensuring their respective places [100]. At the genomic level the two species have proved remarkably reinforcing, highlighting gene structure [61], such as the determination of alternative transcripts [119], chromosomal changes through analysing altered synteny, and the analysis of conserved non-coding regions, the focus of the present study. *C. briggsae* is currently undergoing non-systematic sequencing of particular genomic regions, whereby selected BACs and fosmids built into a fingerprint map are chosen based on hybridization experiments on an investigator requested basis (J. Schein, personal communication). There is a strong desire, once funds are available in the large sequencing facilities of the Sanger Centre and Washington University, to ramp this up to a high-throughput production (M. Marra, personal communication). An approach to detect orthologues between the *C. briggsae* finished sequences housed at Washington University [131] and the *C. elegans* annotated gene collection is outlined in Appendix A.3. From the ensuing set I shall point out two cases.

Highlighting the value of such sequence comparison is Figure 1.6 depicting, in Dotter format [108], the upstream region of predicted gene B0228.3 and its possible orthologue in *C. briggsae.* The sequence is an exon (D. Baillie, personal communication) and therefore an unannotated exon or partial exon of the gene, B0228.3, that was used to find it.

Figure 1.7 depicts the alignment of predicted gene c47d12.3 and its orthologue, their representation in Dotter format [108] and a positional distribution of the amount of cross alignment in terms of potential binding sites in which a match scores 1, a mismatch -1, as described in Appendix A.2. This requires more work to interpret. Indeed, it points

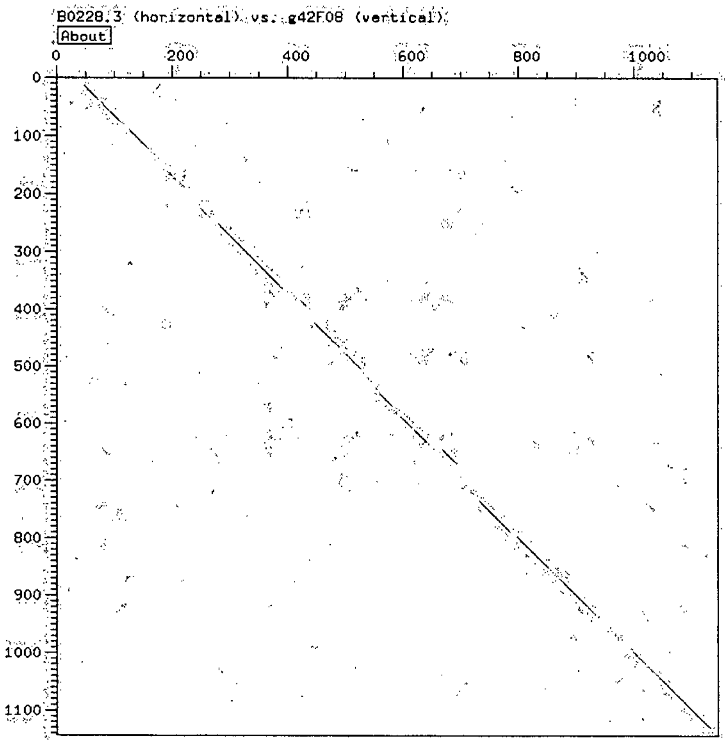


Figure 1.6: **Alignment of the upstream region of predicted gene B0228.3 with its possible orthologue in *C. briggsae* represented in Dotter format [108].**

to a complicated scenario for automation to solve. The difficulty resides in the degree of cross alignment, making any extraction of a local orthology over any other suspect. This hints, whether correct in this case or not, at an approach to extract orthologous non-coding elements. In the modified Smith-Waterman algorithm described previously in Appendix A.2 only the diagonal elements are represented, horizontal and vertical entries having been zeroed out. In extracting the entries once they are scored we might keep tally of the indexes for both the x and y coordinates, disallowing either index to fall below the maximum value achieved at any given point. Considering all possible paths over the traceback procedure, the elements in the path scoring the maximum would be chosen. In many respects this approach is

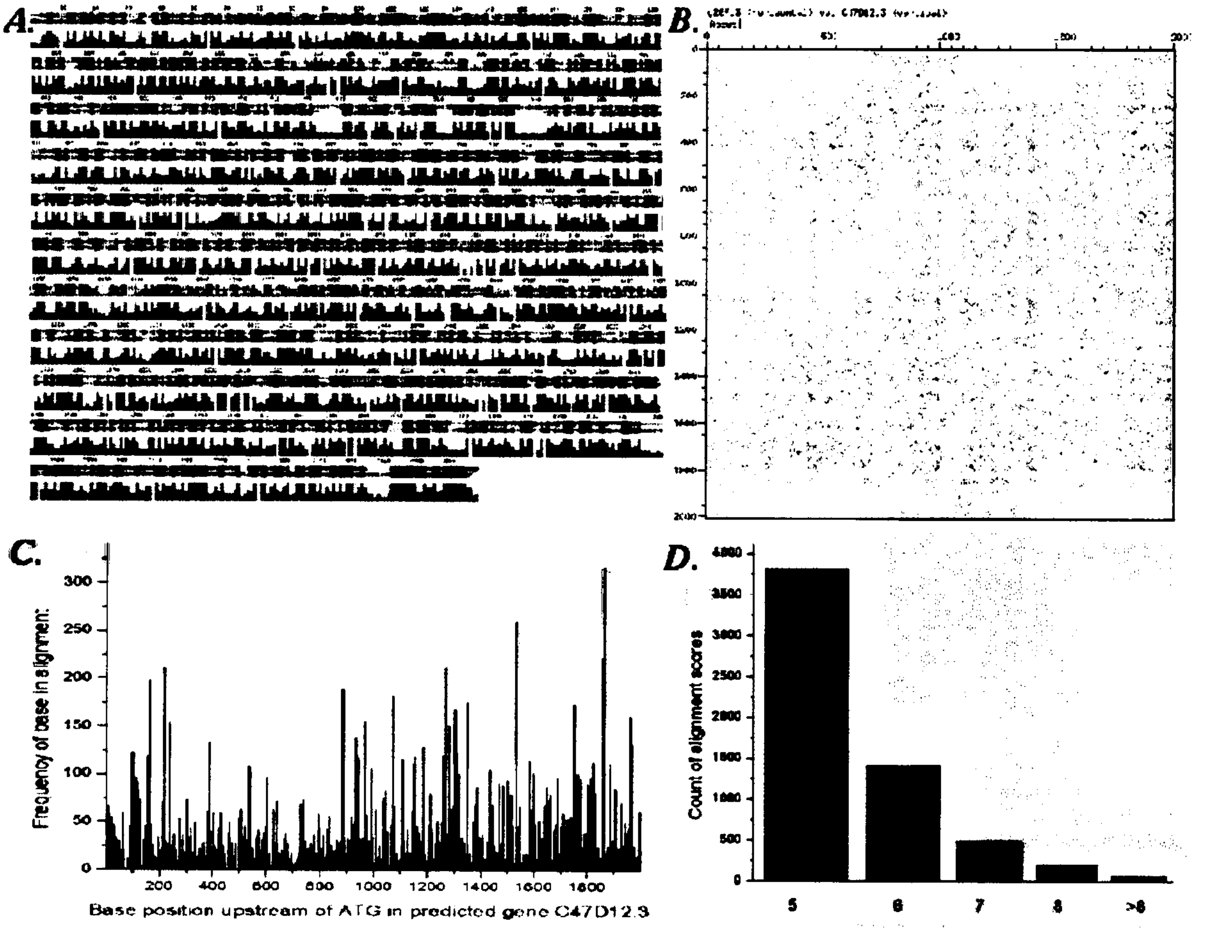


Figure 1.7: **Alignment of the upstream region of predicted gene c47d12.3 with its *C. briggsae* orthologue (A), the representation of the resulting noise in both Dotter (B) and linear (C) format, and a distribution of the scores (D) for aligning the binding sites as described in Appendix A.2.**

akin to treating the Smith-Waterman algorithm in the framework of the Needleman-Wunsch algorithm. That is, the global alignment of the local alignments.

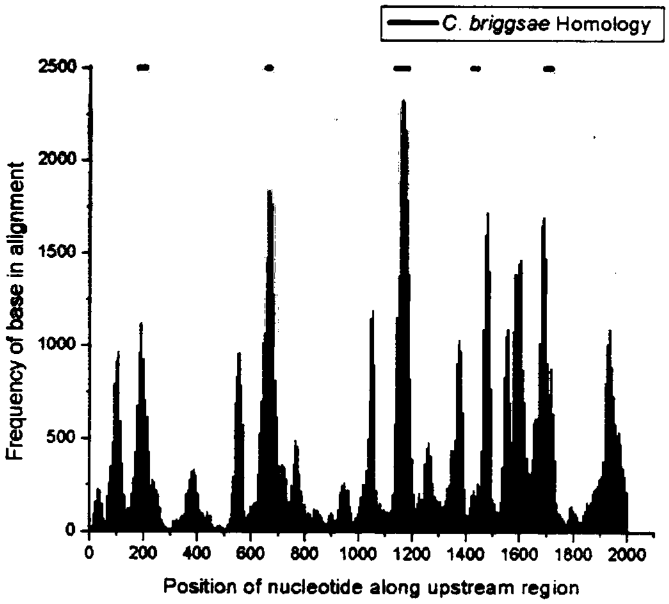


Figure 1.8: **Distribution of the base pairs of the upstream region of predicted gene B0365.6 in alignment with all other upstream regions in the dataset and its *C. briggsae* orthologue.**

A final point of quizzical interest concerning the predicted gene B0365.6 and its *C. briggsae* orthologue. When treated in the manner described previously with K04F10.4e aligned for binding sites against the other 15,524 upstream regions, an extraordinary correlation between the peaks of common alignment and the regions it shares with its *C. briggsae* orthologue can be seen in Figure 1.8.

Chapter 2

Coexpression and Coregulation

The recent technologies for measuring genome-wide expression have created a much deserved excitement in much the same manner as having the full sequence of an organism. And indeed it will be the refinement and advancement of such technologies that shall be the overriding concern in the years to come. It is clear why. They provide a glimpse into the dynamic behaviour of the genes at any order of magnitude that is technologically feasible, from individual cells to tissues to whole multicellular organisms. Consequently, they are irrevocably changing the face of both fundamental biology and clinical medicine [4]. A fascinating aspect of these technologies is that they are drawing on and motivating developments across a wide range of fields, from cell sorting to hybridization techniques to statistical measurement to cluster analysis. There is a great deal of lively debate currently concerning the different approaches to measurement of gene expression with some camps prefering the microarray or gene chip technologies [103], whose strength lie in the ability to easily measure samplings from a number of different time points or developmental stages, with the other camp prefering a more exacting analysis for a given time point or developmental stage, such as with the

SAGE protocol [127]. Both approaches have been applied to *C. elegans.* This study looks briefly at some results from the Stanford Microarray Database [110].

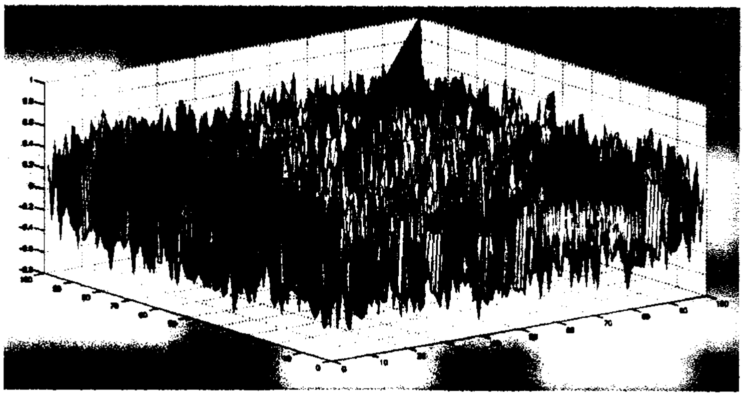


Figure 2.1: **Unclustered correlation surface map of gene expression. The values in the array defining the surface are the pairwise Pearson correlation coefficients over the ratio values from a number of microarray experiments between all genes in the set.**

As stated, among the problems that are being tackled in a number of ways is that of clustering [32] [5] [117]. In its most general form the problem can be seen as a noisy surface of a correlation matrix representing the expression of a set of genes over a number of different experiments or time points as shown in Figure 2.1. In this context, the process of clustering can be separated into two problems, the first is the smoothing of the surface, the points of interest being the global as well as, within a cutoff, the local maxima and minima. The other is in determining the cliques that emerge from the correlation matrix [15].

The methods used in studying gene expression rely on the sampling of relative levels of mRNA product. But mRNA undergoes regulatory control of its own, not only in the translation to its final product but in its life span, each mRNA having potentially different rates of degradation. And even the same mRNA has been shown to have variable decay rates that differ by more than two orders of magnitude [104]. Expression technologies, in order that the information be meaningful in terms of the final product, are interpreted with the implicit assumption of a consistent relationship between translation and mRNA turnover, as well as the role of RNA surveillance for the ensuring of no superfluous mRNA product. There are, however, a number of potential pitfalls with these assumptions in regards to the technologies as they are currently implemented. The existence of polycistronic transcripts in *C. elegans*, while themselves a subject of debate, provide an interesting context in which to examine expression technologies. Clearly, if any mRNA should have similar expression, then these transcripts should. If they don't, and they are truly polycistronic, then mRNA degadation should be the cause, although over a number of time points one might expect them to remain highly correlated. Whether this is the case or not leads to the approach possibly being of value in bringing to light different translational signals or, conversely, as a method for examining expression technologies. Table 2.1 shows a set of polycistronic transcripts described by Zorio *et al* [140].

Operon set	Correlation
mai-1(K10B3.9), *gpd-2*(K10B3.8), *gpd-3*(K10B3.7)	0.510
lin-15b(ZK662.4), *lin-15a*(ZK678.1)	0.314
kin-16(M176.7), *kin-15*(M176.6)	-0.055
ZK353.8, ZK353.7	0.768
C50C3.8, C50C3.7	0.401
ZK637.9, ZK637.10	0.383
ZK637.3, ZK637.5	0.853
R05D3.2, R05D3.1, R05D3.11	0.783
K06H7.4, K06H7.3	0.636
C06E1.10, C06E1.9	-0.054

Table 2.1: **Operons and their clustering correlation from expression data. See Appendix A.4 for details.**

The correlation results in Table 2.1 are derived from a microarray experiment containing genomic PCR products representing 11,917 predicted genes [98] (see Appendix A.4). The results vary widely. Of the two sets of operons that contain three genes i.e. *mai-1*(K10B3.9), *gpd-2*(K10B3.8), *gpd-3*(K10B3.7) and R05D3.1, R05D3.2, R05D3.11, two of the genes correlate a great deal more highly than with the inclusion of the third. For instance, *gpd-1* and *gpd-3*, two functionally related genes, correlate to a high value of 0.927 but with the inclusion of *mai-1* the correlation drops to 0.510. If this is not an artifact of measurement, it is intriguing. It leads to a question of whether *mai-1*, the first upstream gene of the three, is not in fact a seperately transcribed gene, or whether it has different methods of degradation. If it is an artifact of measurement, here lies fuel for the critics. It would be of interest to tackle this approach in a bacterial system, where clearly understood operons are treated.

B asymmetry	*lin-44 lin-17 vab-3*	Emmons and Sternberg, 1997
embryo polarity	*mex-1 par-1 par-2 par-3 par-4 par-5 par-6*	Kemphues and Strome, 1997
muscle actin	*act-1 act-2 act-3 act-4*	Moerman and Fire, 1997
cuticle L3	*sqt-1 rol-6 col-12 col-13 col-1 col-17 col-15*	Moerman and Fire, 1997
cuticle Ld2	*col-1 col-15 col-2 col-6 col-8 col-36 col-40*	Moerman and Fire, 1997
VPC generation	*lin-26 lin-39 unc-83 unc-84*	Greenwald, 1997
inductive signaling	*let-23 let-60 lin-3 lin-45 mek-2 mpk-1 sem-5*	Greenwald, 1997
inhibitory signaling	*lin-9 lin-13 lin-15 lin-36 lin-37*	Greenwald, 1997
vulval fates	*lin-11 lin-17 lin-18 vex-1*	Greenwald, 1997
phagocytosis	*ced-1 ced-6 ced-7 ced-2 ced-5 ced-10*	Hengartner, 1997
migration	*unc-5 unc-6 unc-40*	Antebi *et al*, 1997
major sperm protein	msp family	L'Hernault, 1997
pharyngeal	*egl-2 egl-23 egl-36 exp-3 exp-4 unc-93 sup-10 egl-30 unc-103 unc-58 unc-90 unc-105 unc-43 unc-110*	Avery and Thomas, 1997
cilium structure	*che-2 che-3 che-10 che-11 che-13 daf-10 daf-13 osm-1 osm-5 osm-6*	Bargmann and Mori, 1997
axon guidance	*tax-2 tax-4 daf-11 daf-21*	Bargmann and Mori, 1997
water-soluble chemotaxis	*che-1 che-6 che-15 che-16*	Bargmann and Mori, 1997
osmotic avoidance	*osm-7 osm-8 osm-11 osm-12*	Bargmann and Mori, 1997
HSN path selection	*enu-1 fax-1 unc-42 unc-115*	Antebi *et al*, 1997
daf-c	*daf-1 daf-4 daf-7 daf-8 daf-14*	Riddle and Albert, 1997

Table 2.2: **A sampling of functionally related genes potentially coexpressed through similar transcriptional programs.**

A major premise behind common transcriptional programs is their utility in coordinating tissue specific or developmentally coordinated processes [3]. Table 2.2 shows a collection of genes that are involved in a number of different functional processes in *C. elegans*. From this collection the major sperm protein family provides an interesting set to examine.

2.1 The Major Sperm Protein Family

The major sperm protein (MSP) family consists of roughly 40 small intronless genes dispersed at three chromosomal loci and are exclusively expressed in late primary spermatocytes, where they comprise ≈ 15% of all the protein, playing a role in sperm motility, oocyte maturation and gonadal sheath cell contraction [65] [82]. In this respect they are a unique multigene family in having preserved strict cellular and developmental regulation of expression without being organized in tandem arrays [130]. From the MSP genes annotated in the *C. elegans* database 23 were found in the previously constructed dataset of upstream regions. The recent microarray experiments of Reinke *et al* [98] include 17 of these 23 genes, 10 of which

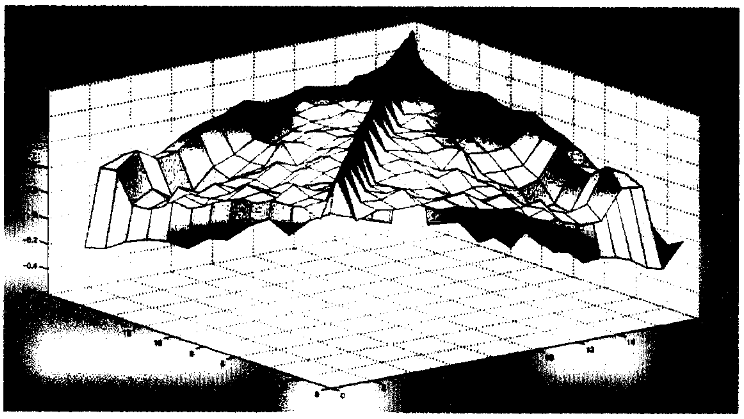


Figure 2.2: **Correlation surface map of 18 MSP genes generated by pairwise Pearson correlation coefficients over all experiments housed in the Stanford Microarray Database [110].**

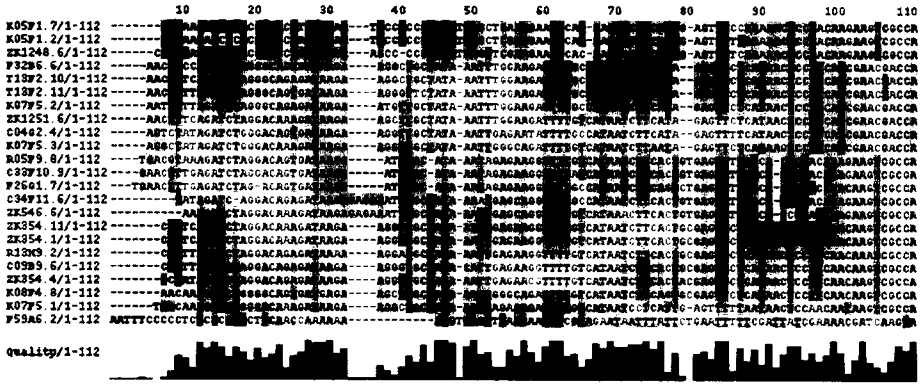


Figure 2.3: **Alignment of ≈ 100*bp* of the upstream region immediately 5′ of the ATG of 22 MSP genes. Aligned with the aid of the ClustalW server at EBI [23].**

were of sufficiently good hybridization quality. These ten cluster together with a correlation of 0.986%, as measured by the software tool Cluster [31]. Indeed, querying all microarray experiments at the Stanford Microarray Database yielded 18 of this set of 23 genes, their strong overall correlation shown in the surface map of Figure 2.2. The degree of overall homology up to ≈ 100*bp* of the upstream regions is striking amongst the 23 MSP genes, as was noted by Klass and Ammons [66] (see Figure 2.3), as well as their degree of cross alignment (see Figure 2.4). A homology search with the upstream region of just one of the MSP genes against the dataset of 15,525 upstream regions pulled out all the other 22 MSP genes in the set, as well as one seven-pass transmembrane chemoreceptor gene, a member of the serpentine receptor class d (srd) multigene family. This high degree of homology in the

Motif	MSP Family	Total	P value
CATAATCTTTCA	16/23	37/15,525	< 0.0001
AGATCT	21/23	4231/15,525	< 0.0001
GATAAGA	16/23	2091/15,525	< 0.0001
TTGCTATAAATT	9/23	20/15,525	< 0.0001

Table 2.3: **Sampling of motifs found in the first 100 bp upstream region of the MSP family and their relative abundance compared with the total dataset.**

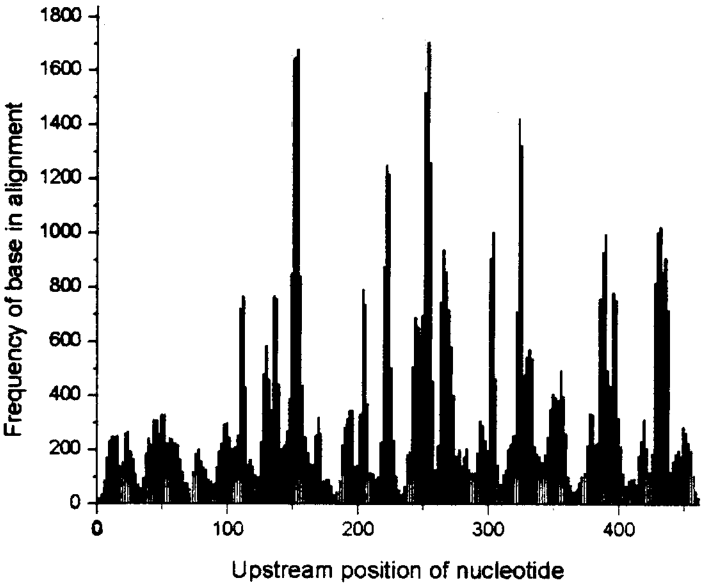


Figure 2.4: **Signal profile of predicted gene C09B9.6 aligned for potential binding sites against nine other MSP genes, generated through the method described in Appendix A.2.**

upstream regions of the MSP set indicates a high likelihood of their having arisen through gene duplication. That said, it would seem a difficult task to extract those elements that are functionally important. Klass and Ammons [66] pointed out, almost arbitrarily, a few conserved motifs from the set they analyzed. Table 2.3 examines these motifs in the light of the set of the 23 MSP genes vis-a-vis the dataset. As can be seen, all were significant under a Fisher exact test. Since the smallest motif was a hexamer this leads to a possibly interesting method for extracting significant motifs (see Appendix A.5). Table 2.4 shows the results, 10 of which prove significant, the hexamer from the Klass and Ammons paper included.

Motif	num	Total num	P value
AACTCC	17	4423	0.0033
AAGAAG	18	7127	0.0639
AATCTT	18	8556	0.1692
ACTCCT	15	3414	0.0016
AGAAGG	14	3967	0.0109
AGATAA	13	6341	0.2217
AGATCT	16	4231	0.0045
ATAAAT	18	11515	0.4919
ATAACT	13	6700	0.2702
ATAAGA	15	6231	0.0998
ATAATC	18	6524	0.0365
ATCTTC	17	7130	0.0938
CATAAT	18	6963	0.0554
CCTTCA	13	4768	0.061
CTATAA	15	5735	0.0642
GATAAG	18	4422	0.0016
GCTATA	15	2525	<0.0001
TAAATT	15	12551	0.7906
TAACTC	14	3881	0.0094
TAAGAA	13	7356	0.3638
TAATCT	17	5933	0.0309
TATAAA	16	10568	0.5343
TCATAA	19	7932	0.0827
TCCTTC	13	5806	0.156
TCTTCA	19	8907	0.1537
TGCTAT	15	3465	0.0018
TTCTCA	14	9880	0.6132

Table 2.4: **Statistical significance, determined by the Fisher exact test, of those hexamers arising from the first principal component in the matrix derived from the 23 MSP genes.**

2.2 The Germ Line

The totipotent and immortal hermaphroditic germline in *C. elegans* is a complex syncitial tissue that undergoes a number of fundamental processes such as spermatogenesis, oogenesis, sex-determination, meiosis, genetic recombination and chromosome re-assortment [64] [102] [105]. One of the first experiments to emerge subjecting *C. elegans* to genome-wide expression technologies, the previously mentioned results from Reinke *et al* [98], studied the genetic changes taking place in the germline. Having built a microarray incorporating genomic PCR products representing 11,917 genes, this group exploited some well known mutants to arrive

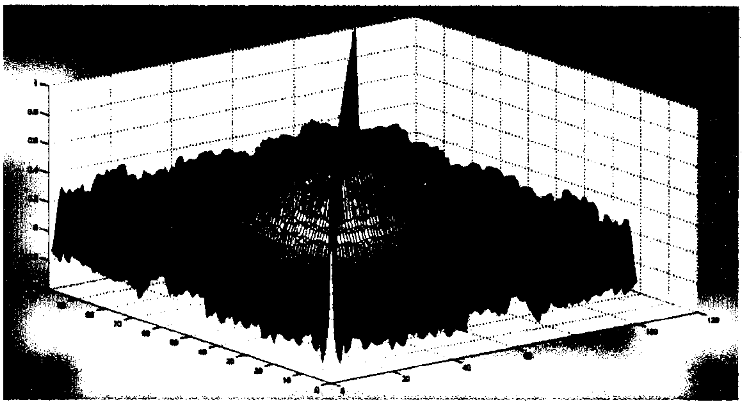


Figure 2.5: **Cross section of the correlation surface map of the germline intrinsic genes, generated by representing as a surface the array of pairwise correlations of microarray expression ratios arising from the set of germline intrinsic genes.**

at a collection of genes that were categorized as being sperm-enriched, oocyte-enriched or germline intrinsic, the latter simply those not particular to the preceding differentiations. It is the 508 germline intrinsic genes with which I shall be concerned. As shown in Figure 2.6, the chromosomal distribution of the germline intrinsic set is similar to what would be expected of genes involved in fundamental and strongly conserved cellular functions.

The reason for examining the germline intrinsic set of genes is for precisely the opposite reason in looking at the MSP set. It would be hoped that a complex patterning of regulatory signatures would emerge from the various nodes that the genes cluster into. However, even with such a complex set the correlations that arise from the microarray data (Figure 2.7) underlie the fact that many of these processes are active at temporally similar occasions and that the populations of worms involved in the preparation of mRNA do not lend themselves to the sort of fine scale expression studies that might be desired in this situation. Applying a hierarchical clustering program [31], the 508 genes, of which 377 are represented in the

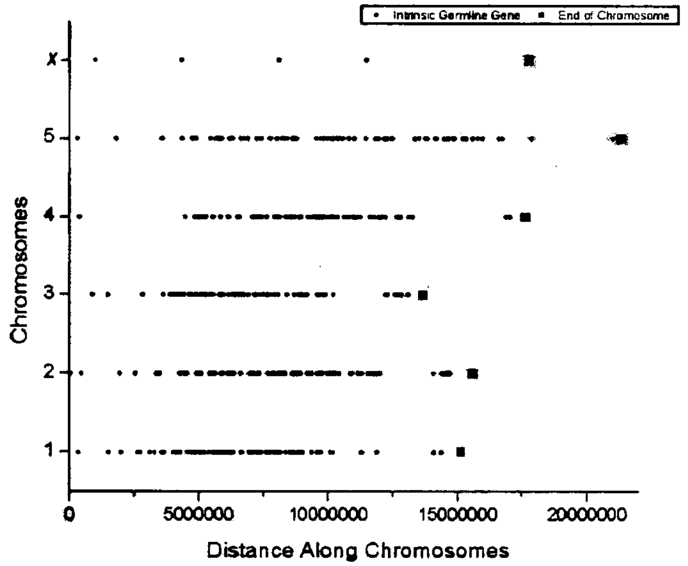


Figure 2.6: **Distribution of germline intrinsic genes over the six chromosomes.**

dataset, break down into 70 nodes. Taking the largest node of 362 genes, and applying the same technique as with the MSP family (see Appendix A.5), building a matrix of the represented hexamers, the first principal component was determined and by the χ^2 test, with the traditional α value of 0.05, all those elements not deemed significant were discarded. The significant hexamers were then positionally ordered according to their place in the upstream regions of the genes comprising the node and pairwise aligned through a modification of the Needleman-Wunsch algorithm (see Appendix A.6). Figure 2.8 depicts the frequency of the runs of three consecutive motifs as a surface map, the peaks being the hexamer signatures for the node. Although to be significant it might be assumed that a particular run of motifs is at least as strongly represented as the number of pairwise alignments resulting from the genes in a node this is clearly not the case in the node just examined. Plotting the maximum run of three motifs divided by the number of pairwise alignments for the entire set of 70

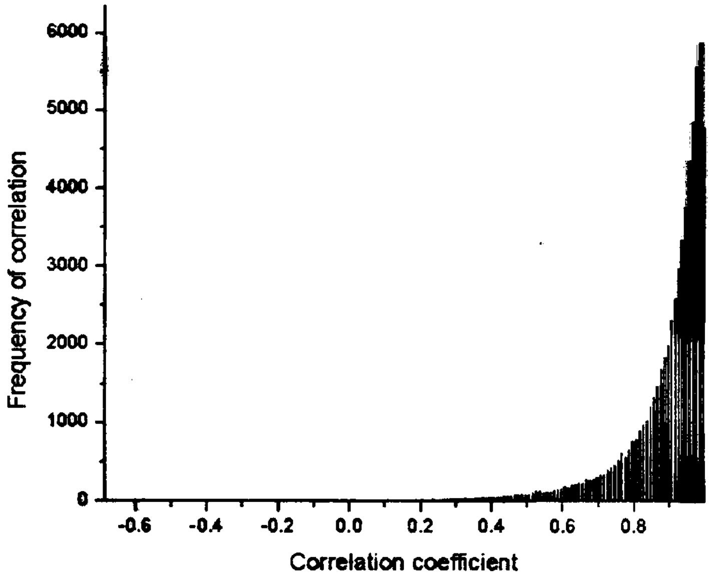


Figure 2.7: **The distribution of the correlation coefficients amongst the nodes clustering the germline intrinsic genes. The correlations were gathered from the nodes generated by the program Cluster [31].**

nodes of the germline intrinsic set results in Figure 2.9. Not a substantial result and indeed shows little that might be construed as a strong argument in favour of the idea of regulatory signatures emerging from microarray clusters. As was mentioned, owing to the experimental protocol, the problematic aspect of this set of genes was the tight correlation of all the nodes. The mechanism of hierarchical clustering did not have to work too hard and the genes may have been found to settle into a very different and finer arrangement with another clustering method, in which case a new set of regulatory signatures might well prove more significant.

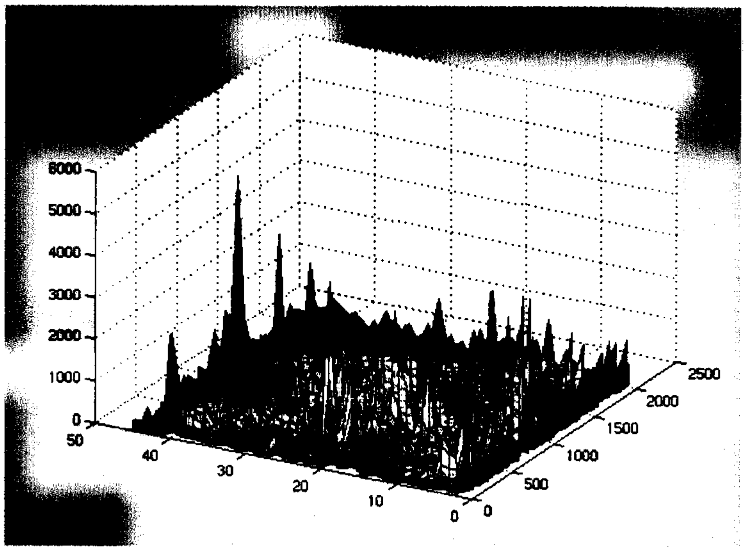


Figure 2.8: **Regulatory signatures emerging from the surface map of the frequencies of all runs of the combinations of three significant motifs. See Appendix A.6 for further details.**

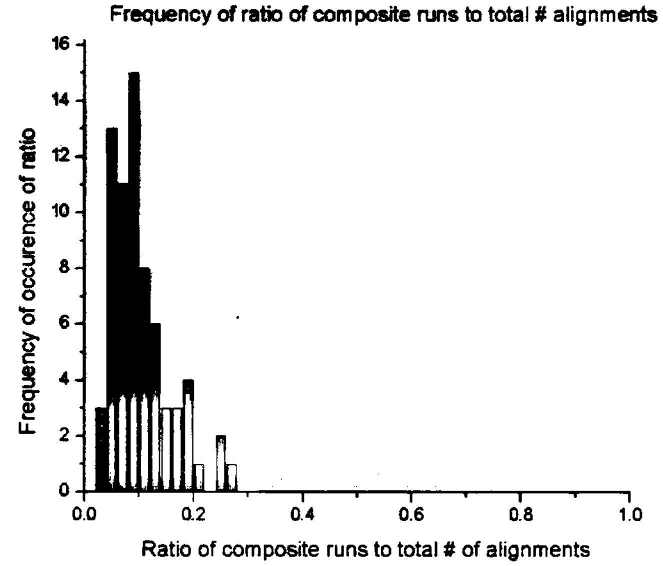


Figure 2.9: **No significant signatures emerge over the set of germline intrinsic genes. The graph was generated by plotting, for all 70 nodes, the number of runs for the maximum group of three significant motifs divided by the total number of pairwise alignments for the node.**

Chapter 3

Transcriptional Profiles

The techniques from the preceding discussions, both of overrepresented oligonucleotides and phylogenetic footprinting, yielded a number of motifs, some of which one hopes have functional relevance, others naturally more suspect. Regardless, they provide a reasonable harvest. From Figure 1.3(B) the red line at the 95% mark is the filter above which all motifs were selected, a total of 1539, many of the smaller motifs embedded in the larger motifs. The gathering of potential elements through phylogenetic footprinting entailed two approaches. The first, which was already discussed in section 1.1, resulted in a set of 163 potential binding sites. A subset of the sequenced BACs and fosmids have also been analysed for orthologous pairs by a team in University of California Santa Cruz [61]. Gathering from their results 1003 orthologous regions upstream of the 5′ UTR in 132 genetically characterized genes, and parsing these for motifs less that 50 bp in length, a set of 554 potential elements were extracted.

This set of motifs were then grouped together into closely associated sequences in order to build motifs into either regular expressions or position weight matrices (PWM's). This

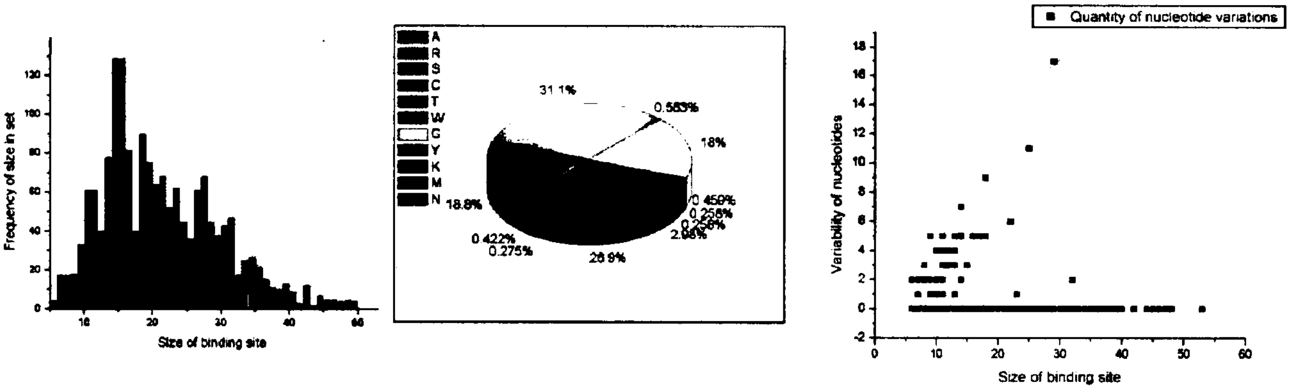


Figure 3.1: **Size distribution, nucleotide composition, and nucleotide variability of binding sites incorporated into the transcriptional profile matrix. See Appendix B for interpreting the IUPAC ambiguity codes.**

procedure is deceptively simple, and is actually the most difficult step in cataloguing computationally derived binding sites. The difficulty stems not simply from the watering down of a motif, the allowance of too much nucleotide variability, but in melding possibly true binding sites together. The complementary fear of being too stringent would entail having to group similar motifs after the fact, based on expression studies. This may well prove the sounder of the two approaches. For this study a simple rule enforced the grouping of two sequences if their sizes were similar and their variability under a certain number. As the size of the binding sites grow this allowance for variability increases. See Appendix A.7 for details surrounding the generation of the 1362 PWMs from the set of overrepresented motifs and phylogenetic footprint results, to which were added 55 PWMs experimental and phylogenetic footprints from the literature and 550 binding sites gathered from the Transfac database [137]. The distribution of sizes, and both the ratio and variability of nucleotides for all the incorporated motifs can be seen in Figure 3.1.

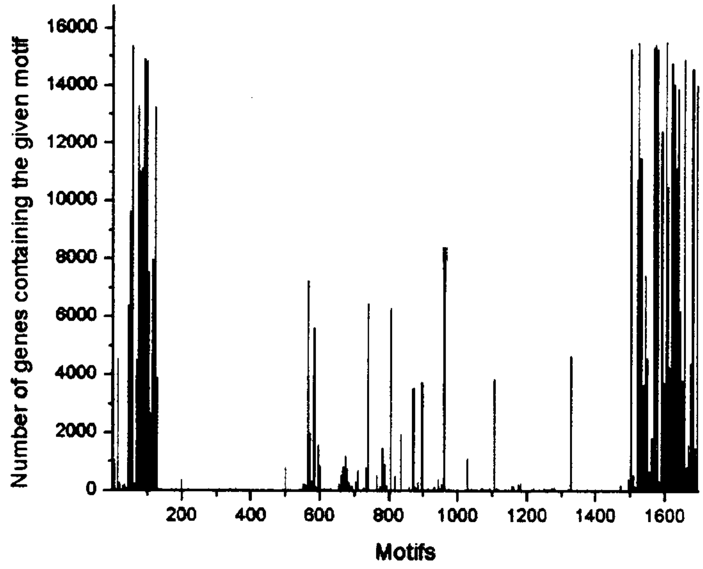


Figure 3.2: **Distribution of the numbers of motifs found in the set of genes.**

The emergent picture of the resulting transcriptional profile matrix is of a small number of elements being shared by a large number of genes (see Figure 3.2 and Table 3.1).

Although such a matrix is a simplification its value rests in the amount of information that can be represented. An extension to the current matrix would be to include any translational signals, and any other contributing factors that may be treated discretely, influential to mRNA expression. An equivalent way to perceive the matrix is as a graph. And in fact this is a more natural definition. In this manner examining the cliques that emerge from the graph is an obvious approach for defining subgroups of genes or regulatory elements, aided by transforming the matrix into adjacency matrices such as in Figure 3.3. Another immediate study would be to examine the adjacency matrices for their connectivity. While I have not pursued this, it is not difficult to imagine the graphs exhibiting the properties of the "small world" hypothesis [135] [113], phenomena that lie in a margin between order and randomness, seen in widely disparate contexts, from social networks to networks derived

Motif	Transfac	Name	Total num	Re value
RTCAT			15442	2.48
TRTTKRYTYS			15506	0.24
NGGRGN	ADR1 01	alcohol dehydrogenase gene regulator 1	15439	12.30
NCRTGTNNWN	MATALPHA2 01	mating factor alpha2	15407	26.82
NNNNNNATTAMYNNNN	DFD 01	Deformed	15474	2.86
SMANAAAAAA	HB 01	Hunchback	15485	2.57
AGAAN	HSF 01	heat shock factor (Drosophila)	15525	29.69
NNNNNNTAATNNNNNNN	UBX 01	Ultrabithorax	15472	15.33
TSTYAMT			15504	0.97
NCANNNNN	CAP 01	cap signal	15525	61.64
MTTTATR	CDXA 01	CdxA	15525	95.09
WWTWMTR	CDXA 02	CdxA	15523	85.62
NNATTRCNNAANNN	CEBPA 01	CCAAT/enhancer binding protein alpha	15506	19.90
RNRTKNNGMAAKNN	CEBPB 01	CCAAT/enhancer binding protein beta	15503	2.07
NNTKTGGWNANNN	CEBP 01	CCAAT/enhancer binding protein	15511	27.08
NNNTTGCNNAANNN	CEBP Q2	CCAAT/enhancer binding factor	15506	21.67
NNNNMGGAWNNNN	CETS1P54 02	c-Ets-1(p54)	15511	8.82
SNNGATNNNN	GATA1 01	GATA-binding factor 1	15522	29.54
NNNGATRNNN	GATA2 01	GATA-binding factor 2	15522	24.28
NNGATARNG	GATA3 01	GATA-binding factor 3	15522	16.86
NNNAAATCANNGNN	GFI1 01	Growth Factor Independence	15442	1.57
NNNAACKGNC	MYB Q6	c-Myb	15524	11.95
NNRTAATNANNN	OCT1 03	octamer factor 1	15524	38.33
AAACWAM	SRY 01	Sex-Determining Region Y	15516	30.95
NWWAACAAWANN	SRY 02	Sex-Determining Region Y	15421	3.94
NNNNNCCATNTWNNNWN	YY1 01	Yin Yang-1	15455	11.61

Table 3.1: **Most represented motifs in the upstream regions of 15,525 genes. The RE value being, for the given motif, the random expectation of a good matrix score in a random sequence of 1000bp [96].**

from metabolic pathways, to the *C. elegans* neuronal pathway [135].

The most fundamental subset of the genes from a regulatory point of view are the transcription factors. As a subset of the genes it is clear that their regulatory logic is quite central to the properties of all other genes under their command. Of the estimated 1300 transcription factors in *C. elegans*, 569 have been associated with predicted genes contained in the dataset, which in turn contains 370 of the 1697 PWM's ($\approx$ 22%) representing the collection of regulatory elements. A clustering of the distances between them (depiction not shown) shows no obvious patterning or subgrouping, an indication that the motifs generated to date remain too indiscriminate.

The properties that emerge from the matrix of an organisms' regulatory infrastructure will be invaluable in building up abstracted pictures and should be of interest when com-

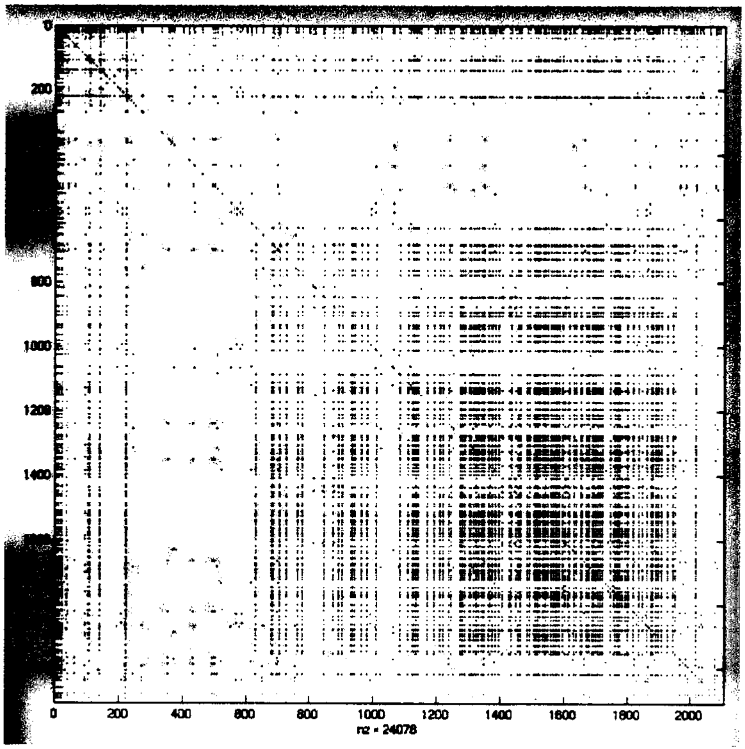


Figure 3.3: **Adjacency matrix of the regulatory elements. An unweighted representation of the Hamming distance between all pairwise columns in the transcriptional profile matrix.**

paring different organisms. But the more immediate problems lie in validating whether our knowledge of the regulatory infrastructure is reasonably correct and, if so, finding methods to associate to each transcription factor or binding site some assessment of its involvement.

3.1 Toward a Deconstruction

Expression data, like the arrays of microarray experiments, may be treated much like a cover on the underlying combinatorics of the transcription factors. I shall make this explicit. Consider two separate matrices. The first, the correlation matrix of pairwise Pearson correlation coefficients of expression data over a set of experiments that was discussed previously, call it $C_{i,j}$. The second, a matrix of the sum of the differences between all pairwise transcriptional vectors. By this is meant the value $\sum |g_i - g_j|$ for transcriptional profile vectors g_i and g_j, maintaining the same index ordering as in $C_{i,j}$. Call this $D_{i,j}$. Then $C_{i,j}$ is, so to speak, a cover over $D_{i,j}$. All those elements of value zero in $D_{i,j}$ conceivably have the same regulatory mechanisms. Their correlations in the matrix $C_{i,j}$ should be within error margin to the value 1. Defining the distribution of the values in $C_{i,j}$ all those elements with zero in $D_{i,j}$ provide statistics on what may be termed the divergence from truth - or falsehood - of the matrix in its containing all regulatory information. So a falsehood measure or index is defined as $1 - F_0$ where

$$F_0 = \frac{\sum C_{i,j}}{\#(D_{i,j} = 0)}, \{\forall i, j | D_{i,j} = 0\},$$

$\#(D_{i,j} = 0)$ the number of such entries. F_0 is an interesting function in that in order for it to snap into place all motifs should be correctly assigned, which is akin to saying individual motifs are held accountable for their missing comrades (I could not use prisoners as this would eventually land me in a prisoners' dilemma). In this respect F_0 has immediate application as a cost function.

Treating the matrix previously constructed to this assessment shows its paucity from calculating the pairwise differences of the transcriptional profiles. Figure 3.4 (left) shows the distribution of a small sampling of these pairwise values, indicating how far we are from a realistic situation. The usual agreement is that eukaryotic genes have, on average, 8 regulatory proteins affecting it [94]. Figure 3.4 (right) shows the same depiction sampling $\approx 70\%$ of the pairwise differences for a matrix constructed previously, containing no motifs

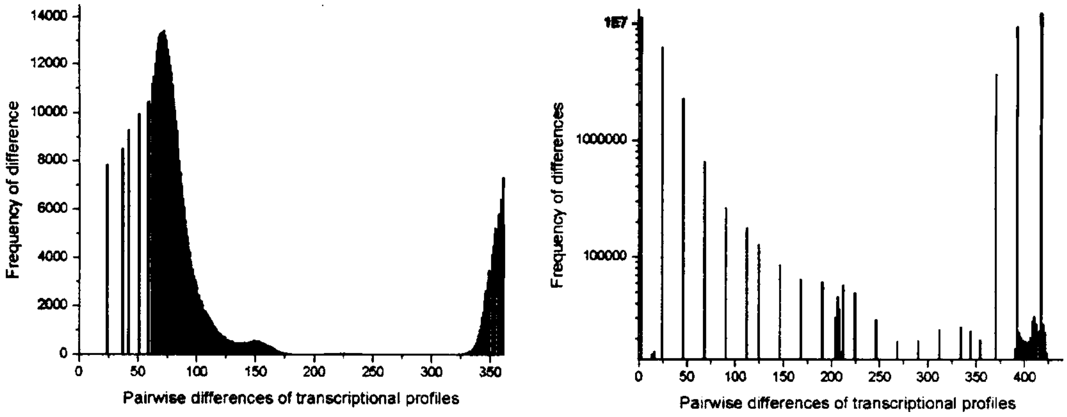


Figure 3.4: **Distribution of the samplings of the pairwise differences between the transcriptional profiles of the matrices generated by PWM's (left) and an exact regular expression approach (right).**

from Transfac or from the experimental literature, and created without use of PWM's but rather exact regular expressions. From this graph there are 3311 pairwise differences of value 0, suitable for investigating the matrix's value for F_0. Using a collection of microarray experiments containing 17,871 genes [53] and averaged over, for each of 6 developmental stages, 2-5 experiments, the values for the corresponding entries in $C_{i,j}$ are plotted in Figure 3.5, providing the atrocious value of $F_0 \approx 0.004$.

When this hurdle is overcome and F_0 brought within error margin to 1 then those values for which $D_{i,j} = 1$ naturally lend themselves to a description and distribution of the influence of a particular motif. In order to effect this a new matrix must be created $E_{i,j}$ which would be sparse except for those entries where $D_{i,j} = 1$ in which case $E_{i,j}$ would take on the value of the position in the vector where g_i and g_j differ, call it k. For each distinct k in $E_{i,j}$ a distribution may be determined by going back to the expression data and plotting, for

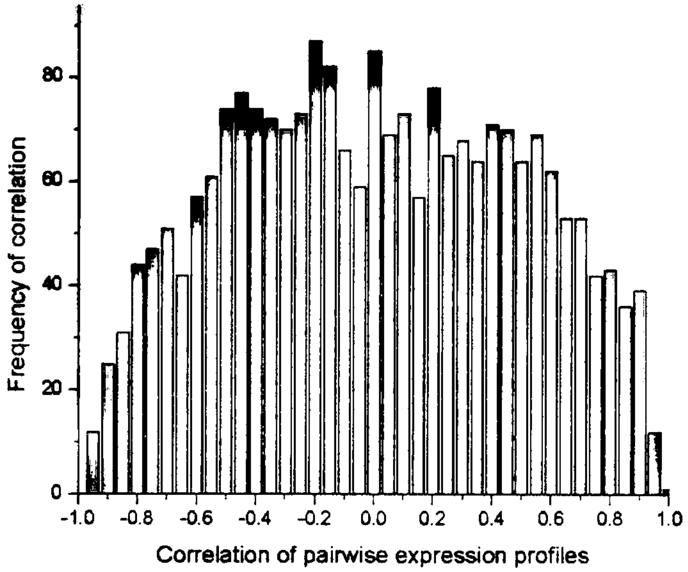


Figure 3.5: **Distribution of the correlation values of an exact matrix with $F_0 \approx 0.004$.**

each set of pairwise genes differing by motif k, and for each time point in the expression matrix, a value of the difference between the expression vectors of the gene containing motif k and the gene without. With all individual motifs of $E_{i,j}$ exhausted in this manner the process naturally extends to generating statistics for a couple and, by increment, groups of transcription factors working together. It would be remarkable if a decoupling were possible between a known motif's statistics and the statistics generated by this motif and another motif whose statistics are unknown. Which could, by recursive descent, extend to one unknown motif and many known motifs. In this manner the complete reverse engineering of regulatory control might very well be within grasp.

3.2 A Measure of Complexity

Complexity is the behaviour emerging from the interaction of aggregate parts. Since transcription is the most central in the hierarchy of cellular processes, a natural definition for the complexity of an organism is implicit in the transcriptional machinery. Treating a transcriptional profile matrix of size m x n, the number of possible unique gene signatures are

$$\sum_{i=1}^{n} \binom{n}{i}$$

which also happens to be the upper bound on the combinations of interacting transcription factors and, therefore, of organismal complexity at this level. Knowledge of the transcription factors and their own signatures gives an even clearer picture of an organisms' complexity. If we were to consider a snapshot of a transcriptional profile matrix of an organism whose gene's profiles are fully worked out then to each regulatory element may be associated a transcription factor, or factors, as in the case of heterodimers. In this respect, we may then consider a given cluster of genes, each of which represent a phase space in the total space of potential clusters, providing a tighter measure of an organisms' complexity. If Ω are the transcription factors $\Omega \subseteq n$, and if σ are the set of regulatory elements of Ω, then

$$\sum_{j=1}^{\sigma} \binom{\sigma}{j} \leq \sum_{i=1}^{n} \binom{n}{i},$$

is an upper bound on the number of phase spaces, and hence a tighter upper bound on the transcriptional complexity of the organism.

Such a measure points to a method to represent the dynamic behaviour of mRNA expression. Since the widespread assumption is that any model of expression would be represented by a set of differential equations it is of interest to see how a first order depiction of the dynamical behaviour of regulation can be viewed in terms of the transcriptional profile matrix. Each phase space of gene clusters is equivalent to the idea of a transcriptome state. And each unique cluster, the result of the action of a set of transcription factors and of a constraint imposed by the inactivity of other present transcription factors, provides a seed for a new round of clustering. Thus, implicit in each phase space is the next phase space, dependent on the reordering based on the columns representing the present transcription factors' binding sites. In this manner the dynamics of the cell may be represented as a continual reshuffling of the transcriptional profile matrix. Indeed, such consideration could well provide a method by which to infer the logic through which transcription factors *en masse* behave.

Conclusion

Unlike Alan Turing's oft quoted statement that "we see a short distance ahead, but we can see plenty there that needs to be done," in the approach of reverse engineering the cell and of model building in this direction we can see both well into the future, as well as an enormous number of immediate hurdles that need be addressed. The study of regulatory systems is many layered: the combinatorics of the transcription factors, their modularity, their altered behaviour dependent on position, their variability in equivalent conditions, epigenetic factors, conformational properties, the vast arena of mRNA expression levels, the dependence on mRNA degradation, posttranscriptional modification.... The list goes on and at every layer the subject remains in its relative infancy, the recent strides of gene expression technologies notwithstanding. Many of the surprises around the corner will, I imagine, be in how information at one layer contributes and informs knowledge in the other layers, as is natural when isolating components of a vast interacting system.

Not surprisingly, the initial aims of this thesis were a great deal broader than they have settled to. It would have been ideal to incorporate as much of the binding site information into the transcriptional profile matrix, including conserved translational regulatory sites and, by association with expression data, deduce the underlying regulatory logic by associating to each binding site (if not transcription factor) a value or mean or even simply a qualitative assessment of its influence. The outrageousness of this scheme is apparent in hindsight only because of greater familiarity with the difficulties. Underlying principles are in effect, however, and abstracting them to create symbolic schemas that may be manipulated in any way to a desired expression level I see as routine, just not immediately.

Ranging over the present work, the clearest statement to be made in its defence is that it is exploratory. No clearly defined goal, apart from the aforementioned, motivated the thesis from its inception. The strongest argument to be made against the work is that a computational method for determining the underlying regulatory logic and, more specifically, the binding sites for transcription factors, is bound to fail - this, because of the enourmous complexity involved and the paucity of our present understanding both *in vivo* and *in vitro*. Even the recent experiments from the Brown and Botstein laboratories [50] [76], describing the binding of purified transcription factors to microarray chips containing intergenic regions from *Saccharomyces cerevisiae*, lead us no furthur in this direction, as interesting as they are. It is trivial to simply pull out base pairs of commonality between any number of segments owing to their correlation by coexpression or phylogenetic relationship. Extending these facts of commonality toward a viable inference of relevance, without aid of a laboratory, is another matter. And in this I have failed. A point which this thesis touches on is that a simple cataloguing of binding sites and their proximal gene is not sufficient evidence for effect. A point well known but rarely displayed. An assumption little questioned in much of the literature on motif searching is the expectation of functional binding sites having more representation in promoter regions proximal to the 5′ UTR. In *C. elegans* the exhaustive study of all n-mers up to 30 shows this to be fallacious. Of what remains in the thesis, little can be validated. A vast number of PWM's have been generated, including a number derived from conserved regions between *C. briggsae* and *C. elegans* that, by all indication, should contain a good quantity of functional sites. All of the analyses based on expression data are crippled somewhat by the crude form of the experiments. They want the finer resolution capable, for instance, with *Saccharomyces cerevisiae.* The methods, for example determining regulatory signatures, lend almost more weight to highlighting the shortcomings of the expression data rather than in isolating relevant information on binding sites. As for the results utilizing the transcriptional profile matrix, a method that can be seen as little more than convenience at present, it must needs prove itself.

As stressed throughout, building expression models of either mRNA or protein or even a combination will, in the final analysis, be made more cohesive by an underlying knowledge of the combinatorics of the transcription factors for each of the genes or gene products. As such, an immediate goal, even if only in a theoretical framework, would be the incorporation of this knowledge, such as with the transcriptional profile matrix, as input into models of interaction.

Model building, whether of protein levels and their interactions, such as with the E-cell [86] and other virtual cell systems, or with mRNA models built from expression changes, have all been valuable attempts but remain precisely that, exploratory tools with experiment as vindicator. There are encouraging signs, however. The growing pains of a subject are usually marked by trial models and simple cases giving way to ever more daring approximations to realistic scenarios. Even though many of the simplest tasks are yet to be overcome, there is an emerging trend of sweeping aside idealized quantities, toward more realistic models stochastic in nature [80] [88]. Not only would these prove more realistic but any attempt to incorporate expression data into the models would be better served.

The true test of all these modelling approaches is whether it is finally accepted by those most practical of individuals, the trained biologists. And the enormity of the task of thinking globally of all the processes in a cell in this more stringent environment is typified by a rather mocking article on the inauguration of Leroy Hood's Institute for Systems Biology in Seattle, Washington [106]. Recently, the Institute delivered its first proof of principle in systems-wide biology, a study perturbing the yeast metabolic pathways [49], incorporating microarrays, quantitative proteomics, and databases resulting in the determination of affected mRNAs, proteins regulated posttranscriptionally and protein-protein interactions. The entire community should be one in hoping that Leroy Hood, and all those of equally audacious outlook, have the last laugh.

Bibliography

[1] Aach, J., *et al.* 2001. Computational comparison of two draft sequences of the human genome. *Nature.* 409:856-859.

[2] Adams, M., *et al.* 2000. The Genome Sequence of *Drosophila melanogaster. Science.* 287:2185-2195.

[3] Alberts, B., *et al.* 1994. *Molecular Biology Of The Cell.* Garland Publishing.

[4] Alizadeh, A., *et al.* 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature.* 403:503-511.

[5] Alter, O., Brown, P., and Botstein, D. 2000. Singular value decomposition for genome-wide expression data processing and modelling. *Proceedings of the National Academy of Sciences USA.* 97:10101-10106.

[6] Altschul, S., *et al.* 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

[7] Andell III, F., *et al.* 1999. Three-Dimensional Structure of the Human TFIID-IIA-IIB Complex. *Science.* 286:2153-2156.

[8] Antebi, A., Norris, C.R., and Hedgecock, E.M. 1997. Cell and Growth Cone Migrations. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[9] Avery, L. and Thomas, J.H. 1997. Feeding and Defecation. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[10] Bargmann, C.I. and Mori, I. 1997. Chemotaxis and Thermotaxis. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[11] Bell, A., West, A., and Felsenfeld, G. 2001. Insulators and Boundaries: Versatile Regulatory Elements in the Eukaryotic Genome. *Science.* 291:447-450.

[12] Blanchette, M. 2001. Algorithms for Phylogenetic Footprinting. *Proceedings of the Fifth Annual International Conference on Computational Biology.* The Association for Computing Machinery.

[13] Blumenthal, T. and Steward, K. 1997. RNA Processing and Gene Structure. *C. ELEGANS II.* Edited by Donald Riddle et al, Cold Spring Harbor Laboratory Press.

[14] Brazma, A. *et al.* 1998. Predicting Gene Regulatory Elements in Silico on a Genomic Scale. *Genome Research.* 8:1202-1215.

[15] Bollobás, B. 1979. *Graph Theory: An Introductory Course.* Springer-Verlag, New York.

[16] Bucher, P. 1999. Regulatory elements and expression profiles. *Curr. Opin. In Struct. Biol.* 9:400-407.

[17] Cartwright, I. and Kelly, S. 1991. Probing the Nature of Chromosomal DNA-Protein Contacts by *In Vivo* Footprinting. *Biotechniques.* Vol. 11, No. 2:188-201.

[18] The *C. elegans* Sequencing Consortium 1991-1998. The *C. elegans* Protein Database Wormpep. `http://www.sanger.ac.uk/Projects/C_elegans/wormpep/`.

[19] The *C. elegans* Sequencing Consortium. 1998. Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. *Science.* 282:2012-2018.

[20] Chen, T., *et al.* 1999. Modelling Gene Expression With Differential Expression. *1999 Pacific Symposium of Biocomputing.*

[21] Clarke, N.D. and Berg, J.M. 1998. Zinc Fingers in *Caenorhabditis elegans*: Finding Families and Probing Pathways. *Science.* 282:2018-2022.

[22] Claverie, J. 2001. What If There Are Only 30,000 Human Genes. *Science.* Vol. 291, No. 5507:1255-1257.

[23] The ClustalW server at EBI. `http://www.ebi.ac.uk/clustalw/`.

[24] Crowley, E., *et al.* 1997. A Statistical Model for Locating Regulatory Regions in Genomic DNA. *J. Mol. Biol.* 268:8-14.

[25] Cornish-Bowden, A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acid Research.* 13:3021-3030.

[26] Davidson, E. 2001. *Genomic Regulatory Systems.* Academic Press.

[27] Desjarlais, J. and Berg, J. 1992. Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proceedings of the National Academy of Sciences USA.* 89:7345-7349.

[28] Durbin, R. and Theirry-Mieg, J. 1989-2001. A *C. elegans* Database. `http://www.acedb.org`.

[29] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G., 1998. *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids.* Cambridge University Press.

[30] Eastman, C., Horvitz, H.R., and Jin, Y. 1999. Coordinated Transcriptional Regulation of the *unc-25* Glutamic Acid Decarboxylase and the *unc-47* GABA Vesicular Transporter by the *Caenorhabditis elegans* UNC-30 Homeodomain Protein. *The Journal of Neuroscience.* vol. 19, No. 15:6225-6234.

[31] Eisen, M. 1998-1999. Cluster and Tree View Manual. preprint.

[32] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA.* 95:14863-14868.

[33] Emmons, S.W. and Sternberg, P.W. 1997. Male Development and Mating Behavior. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[34] Felsenstein, J. 1995. *PHYLIP (Phylogeny Inference Package) Version 3.57c.* `http://evolution.genetics.washington.edu/phylip.html`.

[35] Fickett, J.W. and Hatzzigeorgiou A.G. 1997. Eukaryotic Promoter Recognition. *Genome Research.* 7:861-878.

[36] Fire, A., *et al.* 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans. Nature.* 391:806-811.

[37] Fodor, A., *et al.* 1983. Comparison Of A New Wild-Type *Caenorhabditis Briggsae* With Laboratory Strains of *C. briggsae* and C. elegans. *Nematologica.* 29:203-217.

[38] Genome Sequence Centre, Vancouver, B.C., Canada. `www.bcgsc.bc.ca`.

[39] Gilbert, W. and Muller-Hill, B. 1967. The Lac Operator is DNA. *Proceedings of the National Academy of Sciences USA.* Vol. 58, Issue 6:2415-2421.

[40] Kimball, S. and Mattis, P. GNU Image Manipulation Program, Version1.0.4. `www.gimp.org`.

[41] Gish, W. and States, D. 1993. Identification of protein coding regions by database similarity search. *Nature Genetics.* 3:266-72.

[42] Golub, G. and Van Loan, C. 1996. *Matrix Computations.* Johns Hopkins University Press.

[43] Gower, N.J.D., *et al.* 2001. Dissection of the Promoter Region of the Inositol 1,4,5-Triphosphate Receptor Gene, *itr-1*, in *C. elegans*: A Molecular Basis for Cell-specific Expre ssion of IP_3R Isoforms. preprint.

[44] Greenwald, I. 1997. Development of the Vulva. *C. ELEGANS II.* Edited by Donald Riddle et al, Cold Spring Harbor Laboratory Press.

[45] Hagmann M. 2000. HUMAN GENOME: Mapping a Subtext in Our Genetic Book. *Science.* 288:945b-947b.

[46] Hengartner, M. 1997. Cell Death. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[47] Higgins, D., Thompson, J., and Gibson, T. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* 266:383-402.

[48] Holstege, F., *et al.* 1998. Dissecting the Regulatory Circuitry of a Eukaryotic Genome. *Cell.* 95:717-728.

[49] Ideker, T., *et al.* 2001. Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science.* 292:929-934.

[50] Iyer, V., *et al.* 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature.* 409:533-538.

[51] Jacob, F. and Monod, J. 1961. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *J. Mol. Biol.* 3:318-356.

[52] Jacob, F. 1974. *The Logic of Living Systems: a history of heredity.* Hazell Watson & Viney, Ltd. Aylesbury, Bucks.

[53] Jiang, M., *et al.* 2001. Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans. Proceedings of the National Academy of Sciences USA*. Vol. 98:218-223.

[54] Jacobson, R., *et al.* 2000. Structure and Function of a Human TAFII250 Double Bromodomain Module. *Science.* 288:1422-1425.

[55] Johnson, R. and Baillie, D. 1997. Mutation. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[56] Jones, S. 1999. Computational Analysis of the *Ceanorhabditis elegans* Genome Sequence. Ph.D. Thesis, The Sanger Center.

[57] Karlin, S. and Altschul, S. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences USA*. Vol. 87:2264-2268.

[58] Karlin, S. and Altschul, S. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences USA*. Vol. 90:5873-5877.

[59] Kel-Margoulis, O., *et al.* 2000. COMPEL: A database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Research.* vol. 28, No. 1:311-315.

[60] Kemphues, K.J. and Strome, S. 1997. Fertilization and Establishment of Polarity in the Embryo. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[61] Kent, W.J. and Zahler, A.M. 2000. Conservation, Regulation, Synteny, and Introns in a Large-scale *C. briggsae-C. elegans* Genomic Alignment. 2000. *Genome Research.* 10:1115-1125.

[62] Kent, W.J and Zahler, A.M. 2000. The Intronerator: exploring introns and alternative splicing in *Caenorhabditis elegans. Nucleic Acids Research.* vol. 28, No. 1:91-93.

[63] Kenyon, C. 1997. Environmental Factors and Gene Activities That Influence Life Span. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[64] Kimble, J. and Ward, S. 1988. Germ-line Development and Fertilization. *The Nematode Caenorhabditis elegans.* Edited by William Wood, Cold Spring Harbor Laboratory Press.

[65] Klass, M., Dow, B., and Herndon, M. 1982. Cell-Specific Transcriptional Regulation of the Major Sperm Protein in *Caenorhabditis elegans. Developmental Biology.* 93:152-164.

[66] Klass, M. and Ammons, D. 1988. Conservation in the 5' Flanking Sequences of Transcribed Members of the *Caenorhabditis elegans* Major Sperm Protein Gene Family. *J. Mol. Biol.* 199:15-22.

[67] Kolchanov, N., *et al.* 2000. Transcriptional Regulatory Regions Database (TRRD): its status in 2000. *Nucleic Acids Research.* vol. 28, No. 1:298-301.

[68] Labrador, M., *et al.* 2001. Protein encoding by both DNA strands. *Nature.* 409:1000.

[69] Lamport, L. 1994. *LaTeX: A Document Preparation System.* Digital Equipment Corporation.

[70] Lander, E., *et al.* 2001. Initial Sequencing and analysis of the human genome. *Nature.* 409:860-921.

[71] Latchman, David. 1991. *Eukaryotic Transcription Factors.* Academic Press.

[72] Lee, M., *et al.* 2000. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences USA.* Vol. 97:9834-9839.

[73] Lewin, B. 1997. *Genes VI.* Oxford University Press.

[74] L'Hernault, S.W. 1997. Spermatogenesis. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[75] Lichtsteiner, S. and Tjian, R. 1993. Cloning and properties of the *Caenorhabditis elegans* TATA-box-binding protein. *Proceedings of the National Academy of Sciences USA*. Vol. 90:9673-9677.

[76] Lieb, J., *et al.* 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genetics.* vol. 28:327-334.

[77] Mandel-Gutfreund, Y. and Margalit, H. 1998. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Research.* vol. 26, No. 10:2306-2312.

[78] Matlab. 1984-1999. The MathWorks, Inc.

[79] McAdams, H. and Shapiro, L. 1995. Circuit Simulation of Genetic Networks. *Science.* 269:650-656.

[80] McAdams, H. and Arkin, A. 1997. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences USA*. Vol. 94:814-819.

[81] McGhee, J. and Krause, M. 1997. Transcription Factors and Transcriptional Regulation. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[82] Miller, M., *et al.* 2001. A Sperm Cytoskeletal Protein That Signals Oocyte Meiotic Maturation and Ovulation. *Science.* 291:2144-2147.

[83] Moerman, D.G. and Fire, A. 1997. Muscle: Structure, Function, and Development. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[84] Needleman, S.B. and Wunsch, C.D. 1970. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* 48:443-453.

[85] Nigon, V. and Dougherty, E. 1949. Reproductive Patterns and Attempts at Reciprocal Crossing of Rhabditis Elegans Maupas, 1900, and Rhabditis Briggsae Dougherty and Nigon, 1949. *Journal of Experimental Zoology.* 112:485-503.

[86] Normile, D. 1999. Building Working Cells "in silico". *Science.* Vol. 284, No. 5411:80-81.

[87] Ohler, U. 2000. Promoter Prediction on a Genomic Scale - The *Adh* Experience. *Genome Research.* 10:539-542.

[88] Øksendal, B. 1998. *Stochastic Differential Equations.* Springer Verlag.

[89] *Origin: Data Analysis and Technical Graphics Software.* Microcal Software, Inc.

[90] Pedersen, A. 1998. The Biology of Eukaryotic Promoter Prediction - A Review, preprint.

[91] Périer, R.C., *et al.* 2000. The Eukaryotic Promoter Database (EPD). *Nucleic Acids Research.* Vol. 28, No. 1:302-303.

[92] Pickert L., *et al.* 1998. Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics.* vol. 14, No. 3:244-251.

[93] Prasad, S. and Baillie, D. 1989. Evolutionarily Conserved Coding Sequences in the dpy-20-unc-22 Region of *Caenorhabditis elegans. Genomics.* 5:185-198.

[94] Ptashne, M. 1992. *A Genetic Switch: Phage λ and Higher Organisms.* Cell Press and Blackwell Scientific Publications.

[95] Ptashne, M. and Hopkins, N. 1968. The Operators Controlled by The λ Phage Repressor. *Proceedings of the National Academy of Sciences USA.* Vol. 60, Issue 4:1282-1287.

[96] Quandt, K., *et al.* 1995. MatInd and MatInspector:new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research.* Vol. 23, No. 23:4878-4884.

[97] Red Hat's Linux operating system. `www.redhat.com`.

[98] Reinke, V., *et al.* 2000. A global profile of germ line gene expression in *C. elegans. Molecular Cell.* 6:605-616.

[99] Riddle, D.L and Albert, P.S. 1997. Genetic and Environmental Regulation of Dauer Larva Development. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[100] Riddle, D.L., Blumenthal, T., Meyer, B.J. and Priess, J.R. 1997. Introduction to *C. elegans. C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[101] Ruvkun, G. and Hobert, O. 1998. The Taxonomy of Developmental Control. *Science.* 282:2033-2041.

[102] Schedl, T. 1997. Developmental Genetics of the Germ Line. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[103] Schena, M., *et al.* 1995. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science.* 270:467-470.

[104] Schwartz, D.C. and Parker, R. 2000. Interaction of mRNA Translation and mRNA Degradation in *Saccharomyces cerevisiae. Translational Control of Gene Expression.* Edited by Sonenberg, N., Hershey, J.W.B., and Mathews, M.B., Cold Spring Harbor Laboratory Press.

[105] Seydoux, G. and Schedl, T. 2001. The Germline in *C. elegans*: Origins, Proliferation, and Silencing. *International Review of Cytology.* Vol. 203, Edited by Etkin, L.D. and Jeon, K.W., Academic Press.

[106] Smaglik, P. 2000. For my next trick... *Nature.* 407:828-829.

[107] Smith, T.F. and Waterman, M. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology.* 147:195-197.

[108] Sonnhammer, E. and Durbin, R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene.* 167:GC1-10.

[109] Spieth, J., Brooke, G., Kuersten, S., Lea, K., and Blumenthal, T. 1993. Operons in *C. elegans*: Polycistronic mRNA Precursors Are Processed by Trans-splicing of SL2 to Downstream Coding Regions. *Cell.* 73:521-532.

[110] Stanford Microarray Database. `http://genome-www4.stanford.edu/MicroArray/SMD/`.

[111] Stein, L. and Thierry-Mieg, J. 1998. Scriptable Access to the *Caenorhabditis elegans* Genome Sequence and Other ACEDB Databases. *Genome Reasearch.* 8:1308-1315.

[112] Stormo, G. 1990. Consensus Patterns in DNA. *Methods in Enzymology.* Vol. 183:211-221.

[113] Strogatz, S. 2001. Exploring complex networks. *Nature.* 410:268-276.

[114] Streit A., *et al.* 1999. Homologs of the *Caenorhabditis elegans* Masculinizing Gene *her-1* in *C. briggsae* and the Filarial Parasite *Brugia malayi. Genetics.* 152:1573-1584.

[115] Sulston, J. 1988. Cell Lineage. *The Nematode CAENORHABDITIS ELEGANS.* Edited by William Wood, Cold Spring Harbor Laboratory Press.

[116] Suzuki, M. and Yagi, N. 1994. DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proceedings of the National Academy of Sciences USA.* Vol. 91:12357-12361.

[117] Tamayo, P., *et al.* 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences USA.* 96:2907-2912.

[118] Tan, K., *et al.* 2001. A Comparative Genomics Approach to Prediction of New Members of Regulons. *Genome Reasearch.* 11:566-584.

[119] Thacker, C., Marra, M., Jones, A., Baillie, D. and Rose, A. 1999. Functional Genomics in *Caenorhabditis elegans*: An Approach Involving Comparisons of Sequences from Related Nematodes. *Genome Research.* 9:348-359.

[120] Thieffry D. and Thomas R. 1998. Qualitative analysis of gene networks. *Pac. Symp. Biocomput.* 3:77-88.

[121] Thom, R. 1975. *Structural Stability and Morphogenesis: An Outline of a General Theory of Models.* W.A. Benjamin, Inc.

[122] Thompson, J., Higgins, D. and Gibson, T. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research.* 22:4673-4680.

[123] Tucker, A. 1984. *Applied Combinatorics.* John Wiley & Sons, Inc.

[124] Van Helden *et al.* 1998. Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies. *J. Mol. Biol.* 281:827-842.

[125] Van Helden *et al.* 2000. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research.* Vol. 28, No. 8:1808-1818.

[126] Vatcher, G. 1999. Ph.D. Thesis, Simon Fraser University.

[127] Velculescu, V. *et al.* 1995. Serial Analysis of Gene Expression. *Science.* 270:484-487.

[128] Velculescu, V., *et al.* 1997. Characterization of the Yeast Transcriptome. *Cell.* 88:243.

[129] Wall, L., Christiansen, T., and Schwartz, R. 1996. *Programming Perl.* O'Reilly & Associates, Inc.

[130] Ward, S., *et al.* 1988. Genomic Organisation of Major Sperm Protein Genes and Pseudogenes in The Nematode *Caenorhabditis elegans. J. Mol. Biol.* 199:1-13.

[131] Washington University *C. briggsae* ftp site. `http://genome.wustl.edu:8021/pub/gsc1/sequence/st.louis/briggsae/`.

[132] Washington University *C. briggsae - C. elegans* clone matches. `http://genome.wustl.edu/gsc/Projects/C.briggsae/brig_final.php`.

[133] Wasserman, W. and Fickett, J. 1998. Identification of Regulatory Regions which Confer Muscle-Specific Gene Expression. *J. Mol. Biol.* 278:167-181.

[134] Waterston, R. Sulston, J. and Coulson, A. 1997. The Genome. *C. ELEGANS II.* Edited by Donald Riddle *et al*, Cold Spring Harbor Laboratory Press.

[135] Watts, D. 1999. *Small Worlds: The Dynamics of Networks between Order and Randomness.* Princeton University Press.

[136] Wigner, E. 1960. The Unreasonable Effectiveness of Mathematics in the Natural Sciences. *Communications in Pure and Applied Mathematics.* Vol. 13, Num. 1.

[137] Wingender, E., *et al.* 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Research.* vol 28, No. 1:316-319.

[138] Yanofsky, C. 1992. Transcriptional Regulation: Elegance in Design and Discovery. *Transcriptional Regulation.* Edited by McKnight, S.L. and Yamamoto, K.R., Cold Spring Harbor Laboratory Press.

[139] Yuh C., *et al.* 1998. Genomic Cis-Regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene. *Science.* 279:1896-1902.

[140] Zorio, D.A.R., Cheng, N.N., Blumenthal, T., and Spieth, J. 1994. Operons as a common form of chromosomal organization in *C. elegans. Nature.* 372:270-272.

[141] Zucker-Aprison E. and Blumenthal T. 1989. Potential Regulatory Elements of Nematode Vitellogenin Genes Revealed by Interspecies Sequence Comparison. *J. Mol. Evol.* 28:487-496.

[142] Zuckerkandl, E. and Pauling, L. 1962. Molecular Disease, Evolution, and Genic Heterogeneity. *Horizons in Biochemistry.* Edited by Kasha, M. and Pullman, B. Academic Press. 189-225.

Appendix A

Methodologies

The tools used in this study have been run on a networked system at the Genome Sequence Centre, British Columbia Cancer Research Centre, Vancouver, B.C., Canada [38], consisting of $\approx$ 30 computers, the majority of which are Intel based dual 600 P3's, with 512M RAM, running Red Hat's distribution of Linux [97]. The programs used throughout consist of Perl [129], AcePerl [111], ACEDB [28], MatInd and MatInspector [96], Matlab [78], Blastx [41], Waba [61], ClustalW [47], Dotter [108], Cluster and Treeview [31], and Phylip [34]. So as to leave no stone unturned, the thesis itself was coddled together using Gimp [40], Origin graphing software [89] on the Microsoft Windows operating sytem, and the $\LaTeX$ typesetting system [69].

A.1 Interaction With the *C. elegans* Genome Sequence

Interaction with the *C. elegans* genome sequence has been through the ACEDB [28] database, version WS9, via the AcePerl scripting language [111], which has been limited to the extraction of the upstream regions of all genes. An arbitrary size of 2000bp upstream of the ATG start site of each chosen predicted gene has been used as the maximum cutoff. To counter including possible downstream genes of polycistronic transcripts, regardless of whether binding sites might well reside there, consideration has been taken of a pre-genome assessment of a bimodal distribution tending to have a maximum at 400bp, as well as an argument that 25% of genes might be polycistronically transcribed [13]. Through the above mentioned database all upstream regions of predicted genes (only one when the gene in question had alternative transcripts) were chosen with a cutoff requiring all genes to be at least 700bp from the nearest upstream gene (on the same strand). The distribution of intergenic distances can be seen in Figure A.1.

The cutoffs $>$ 700bp and $<$ 2000bp, from which was excluded, when applicable, 500bp from the next upstream gene resulted in a set of 15,525 gene's upstream regions, $\approx$ 81% of the total (as of this release), and a dataset, ranging from 200-2000bp of nucleotides for each upstream region, $\approx$ 26% of the 97MB total nucleotide content. Comparing the ratio of nucleotides in this set compared to the whole genome shows a fairly consistent representation (Figure A.2).

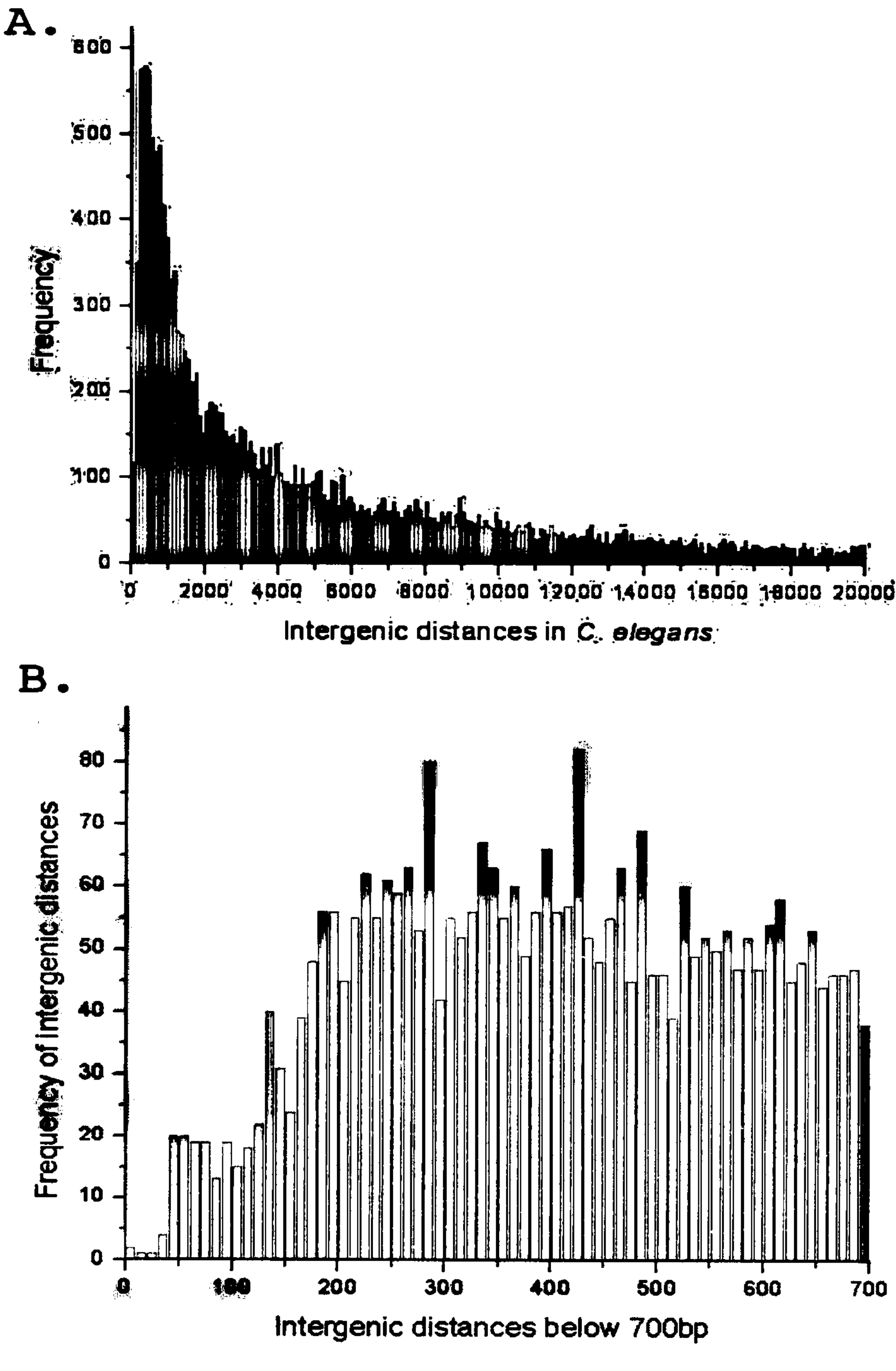


Figure A.1: **Distribution of intergenic distances (A) and those below 700bp (B) in *C. elegans*. The survey used a total of 18,706 intergenic distances with 1923 exceeding the 20,000 mark and 3168 below 700bp.**

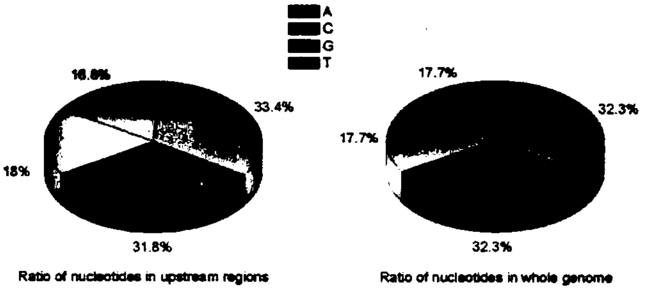


Figure A.2: **Ratio of nucleotides in the selected upstream regions comprising the dataset vs. the whole genome.**

A.2 A Modification of the Smith-Waterman Algorithm for Binding Sites

The Smith-Waterman algorithm is defined, in its simplest form, whereby every gap is assigned a penalty d, as

$$F(i,j) = max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) + d \\ F(i,j-1) + d \\ 0 \end{cases}$$

with s(x_i, y_j) being the scoring of two elements based on a nucleotide distance matrix at position (i,j). In searching for potential binding sites, the consideration of gaps may be waived. The above algorithm can then be rewritten,

$$F(i,j) = max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ 0 \end{cases}$$

When considering dyads or motifs seperated by set spaces, such elements can be found on common diagonals during the traceback of the dynamic programming algorithm. In this respect it might prove interesting to ask whether two identical dimer binding sites are different by one or two base pairs separation, pointing to possibly different helical conformations to be active.

This technique, used in numerous places throughout the present work, makes use of the simplest of scoring mechanisms for s(x_i, y_j): a match is 1, a mismatch -1. In recovering an alignment on the diagonal the default of a minimum of score 5 must be met and the region from the maximum value achieved upward to its drop below the score of 0 is retrieved.

A.3 Phylogenetic Footprinting

An approach to detect orthologues between 511 *C. briggsae* finished sequences representing 12,023,822bp, $\approx$ 12% of the genome, housed at Washington University [131] and the *C. elegans* annotated gene collection entailed their alignment, through the use of Blastx [41], against the database of *C. elegans* translated gene products, Wormpep [18]. The results of this alignment was the detection of the syntenous and orthologous regions in which the *C. elegans* proteins with *C. briggsae* matches shared the same ordering on the *C. elegans* genome as on the *C. briggsae* sequence. When three or more proteins were anchored in this way, the first exon from each of the identified *C. elegans* genes was used to find the orthologous *C. briggsae* exon through an alignment program customized for divergent comparisons [61]. A strict requirement was that the initial alignment had to begin within 50bp of the start of the *C. elegans* exon. This step proved extremely stringent, no doubt too much so. Indeed the entire filtering process brought the potential number of orthologues down from 1356 in the step aligning for synteny to 95 eventual candidates. The stringency was for a reason, however. It was necessary to ensure that the two upstream regions compared were orthologues, using a strictly computational approach, without manual check. A verification of sorts was corroboration from a similar analysis conducted at Washington University [132].

A.4 Microarray Experiments

The microarray experiments used throughout are housed at the Stanford Microarray Database [110]. The particular experiments were those generated for the papers of Reinke *et al* [98] and Jiang *et al* [53], representing genomic PCR products of 11,917 and 17,871 predicted genes respectively. The results were analyzed through the programs Cluster and Treeview, written by Michael Eisen [31], or by grouping the experiments together for each pairwise set of genes by Pearson correlation coefficients and viewing the resulting pairwise gene array through Matlab [78].

The analysis of operons entailed querying the set of experiments generated by Reinke *et al* [98] for the collection of operons described by Zorio *et al* [140] and, through Cluster [31], generate the correlation coefficients.

A.5 Decomposing the MSP Family

By building a matrix containing all hexamers along the columns and the set of genes of interest down the rows and simply marking with a 1 when the upstream region of a gene has the hexamer in question and a 0 otherwise, produces a large sparse matrix. From this matrix, one can determine the most represented motifs through principal component analysis, a method relying on the technique of singular value decomposition. With the hexamers that emerge from the first principal component one can then use the Fisher exact test to ask whether the set of motifs are also significant in relation to the entire dataset, as a method of filtering out those motifs that are more likely to be occuring randomly.

Singular Value Decomposition is a technique that, for the purposes outlined in this thesis, enables the principal components of a correlation matrix to be determined. Taken from Golub and Van Loan [42], pg. 70, the theorem of singular value decomposition is stated as follows: If A is a real m-by-n matrix, then there exist orthogonal matrices

$$U = [u_1, .., u_m] \in \Re^{mxm} \ and \ V = [v_1, .., v_n] \in \Re^{nxn}$$

such that

$$U^T AV = diag(\sigma_1, ..., \sigma_p) \in \Re^{mxn} \ p = min\{m, n\}$$

where

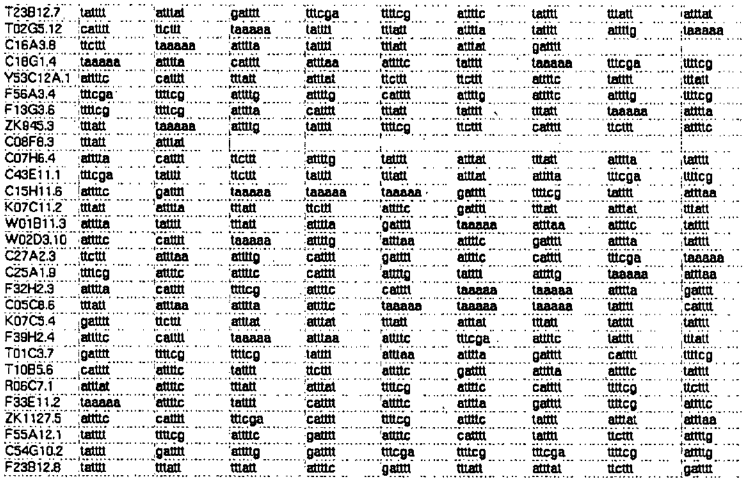
$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p \geq 0.$$

The values of the columns of AV are the principal components. Each principal component (orthogonal to all the other principal components) represents the variability in the dataset according to the value of the eigenvalue associated with it, i.e.

$$\frac{\sigma_i}{trace(U^T AV)}$$

A.6 Regulatory Signatures in the Germ Line

Making use of the Cluster program [31] to generate the correlation coefficients of the hierarchically clustered nodes, each node has been used to generate a matrix of hexamers with 500bp upstream of each gene's translational start site, as described in Appendix A.5. Each of the first principal components of each node were then treated to a χ^2 test, with the traditional α value of 0.05, with all those elements not deemed significant discarded. The significant elements were then positionally ordered according to their place in the upstream regions of the genes comprising the node. Figure A.3 depicts this for clarity. With the genes represented in this manner a modification of the Needleman-Wunsch algorithm [84] may be used. Given two sequences of length n and m, the original algorithm of Needleman-Wunsch used the method of dynamic programming to course through a nxm matrix scoring in each position according to



T23B12.7	tatttt	atttat	gatttt	tttcga	ttttcg	attttc	tatttt	tttatt	atttat
T02G5.12	catttt	ttcttt	taaaaa	tatttt	tttatt	atttta	tatttt	attttg	taaaaa
C16A3.8	ttcttt	taaaaa	atttta	tatttt	tttatt	atttat	gatttt		
C18G1.4	taaaaa	atttta	catttt	atttaa	attttc	tatttt	taaaaa	tttcga	ttttcg
Y53C12A.1	attttc	catttt	tttatt	atttat	ttcttt	ttcttt	attttc	tatttt	tttatt
F56A3.4	tttcga	ttttcg	attttg	attttg	catttt	attttg	attttc	attttg	ttttcg
F13G3.6	ttttcg	ttttcg	atttta	catttt	tttatt	tatttt	tttatt	taaaaa	atttta
ZK845.3	tttatt	taaaaa	attttg	tatttt	ttttcg	ttcttt	catttt	ttcttt	attttc
C08F8.3	tttatt	atttat							
C07H6.4	atttta	catttt	ttcttt	attttg	tatttt	atttat	tttatt	atttta	tatttt
C43E11.1	tttcga	tatttt	ttcttt	tatttt	tttatt	atttat	atttta	tttcga	ttttcg
C15H11.6	attttc	gatttt	taaaaa	taaaaa	taaaaa	taaaaa	gatttt	ttttcg	tatttt	atttaa
K07C11.2	tttatt	atttta	tttatt	ttcttt	attttc	attttc	gatttt	tttatt	atttat	tttatt
W01B11.3	atttta	tatttt	tttatt	atttta	gatttt	gatttt	taaaaa	atttaa	attttc	tatttt
W02D3.10	attttc	catttt	taaaaa	attttg	atttaa	atttaa	attttc	gatttt	atttta	tatttt
C27A2.3	ttcttt	atttaa	attttg	catttt	gatttt	gatttt	attttc	catttt	tttcga	taaaaa
C25A1.9	ttttcg	attttc	attttc	catttt	attttg	attttg	tatttt	attttg	taaaaa	atttaa
F32H2.3	atttta	catttt	ttttcg	attttc	catttt	catttt	taaaaa	taaaaa	atttta	gatttt
C05C8.6	tttatt	atttaa	atttta	attttc	taaaaa	taaaaa	taaaaa	taaaaa	tatttt	catttt
K07C5.4	gatttt	ttcttt	atttat	atttat	tttatt	tttatt	atttat	tttatt	tatttt	tatttt
F39H2.4	attttc	catttt	taaaaa	atttaa	attttc	attttc	tttcga	attttc	tatttt	tttatt
T01C3.7	gatttt	ttttcg	ttttcg	tatttt	atttaa	atttaa	atttta	gatttt	catttt	ttttcg
T10B5.6	catttt	attttc	tatttt	ttcttt	attttc	attttc	gatttt	atttta	attttc	tatttt
R06C7.1	atttat	attttc	tttatt	atttat	ttttcg	ttttcg	attttc	catttt	ttttcg	ttcttt
F33E11.2	taaaaa	attttc	tatttt	catttt	attttc	attttc	atttta	gatttt	ttttcg	attttc
ZK1127.5	attttc	catttt	tttcga	catttt	ttttcg	ttttcg	attttc	tatttt	atttat	atttaa
F55A12.1	tatttt	ttttcg	attttc	gatttt	attttc	attttc	catttt	tatttt	ttcttt	attttg
C54G10.2	tatttt	gatttt	attttg	gatttt	tttcga	tttcga	ttttcg	tttcga	ttttcg	attttg
F23B12.8	tatttt	tttatt	tttatt	attttc	gatttt	gatttt	tttatt	atttat	ttcttt	gatttt

Figure A.3: **Upstream regions of genes in a node represented by the positional placing of significant hexamers.**

$$F(i,j) = max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) + d \\ F(i,j-1) + d \end{cases}$$

s(x_i, y_j) being the scoring of two elements based on a nucleotide distance matrix at position (i,j). The utility of this algorithm for global alignment problems is widespread. For the present purpose the idea is to pairwise align the motifs of significance in the upstream region rather than base pairs over all genes comprising a node. Naturally the possibility exits for extending the algorithm to allow for the alignment of approximate motifs which provides an immediate analogy to the variants in mismatch scoring. This produces a number of alignments depicting the positional similarity of the motifs over the set of genes. From these pairwise alignments of significant motifs the frequency of occurrence of any successive length may be determined for all combinations of significant hexamers. For the largest node in the germline intrinsic set the mean is plotted in Figure A.4 for the number of the given size of run of significant motifs. As a consequence a reasonable size of run for extracting a regulatory signature is 3, rather than the overabundant runs of 2.

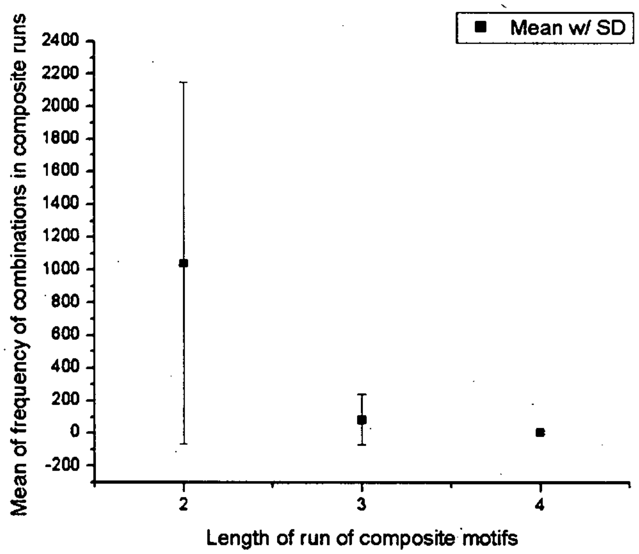


Figure A.4: **Numbers of composite runs of significant motifs in the largest node of the germline intrinsic set.**

A.7 Transcriptional Profiles

	TATAA			WGATAR			RTCAT		
Y48G1A.e	1	1	(747)	1	1	(439)	0	0	0
Y39G10AL.c	1	3	(1909,469,167)	1	2	(811,438)	1	2	(1786,1607)
W03F11.6	0	0	0	0	0	0	1	1	(376)
Y105E8E.f	0	0	0	1	2	(1105,983)	1	4	(1422,1278,443,133)
Y23H5A.5	0	0	0	0	0	0	0	0	0
W09C3.4	1	6	(1368,469,444,312,297,244)	0	0	0	1	2	(678,149)

Table A.1: **The entries of binary, weighted, and positional matrices. See appendix B for interpreting the IUPAC ambiguity codes.**

With the 1539 motifs gathered by their overabundance in the dataset of upstream regions, as well the 717 phylogenetic footprinting results, the next step entailed their being grouped into position weight matrices (PWM). The choice of using PWM's rather than the more constraining approach of regular expressions was decided based on the numerous studies justifying the technique [112]. The rules for the construction of PWM's based on the differing nucleotide composition of the contributing motifs were as follows: below 7 nucleotides the value for allowing a join was 2, from 7 to 10, 3, 10 to 17, 4 and above 17, 5. Two sequences of unequal lengths could be grouped if they differed by no more than 2, in which case the smallest sequence determined the rule by which they could join. The resulting set of 1362 PWM's were constructed from the contributing sequences, the alignments of which were generated by ClustalW [122], with the aid of the program MatInd [96]. To this constructed set were added 55 PWM's generated from experimental and phylogenetic footprints from the literature, as well as 550 binding sites gathered from the Transfac database [137].

With the full complement of all the PWM's, both gathered and generated, it was then necessary to return to the dataset of upstream regions defined in Appendix A.1 and determine which elements lie in each gene's upstream region. To make best use of the PWM's I ran

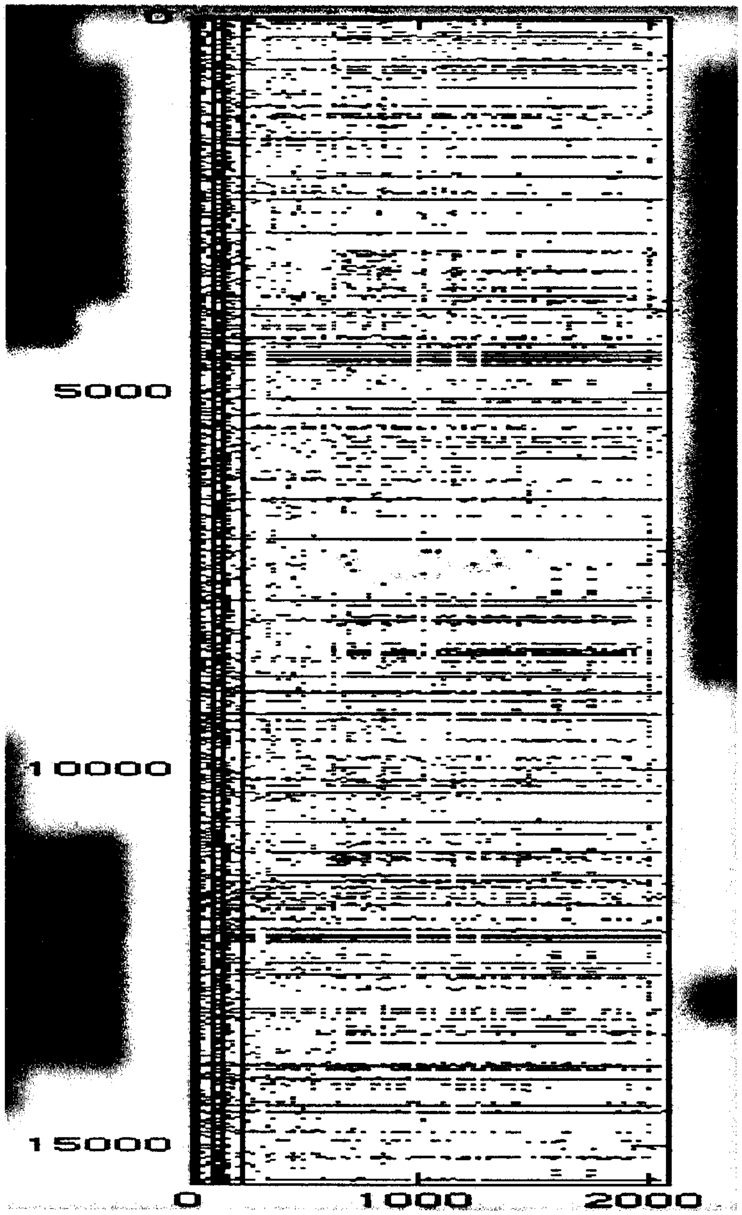


Figure A.5: **Depiction of a binary transcriptional profile matrix in a sparse matrix representation. The rows represent the collection of 15,525 genes in the dataset, while the columns represent the potential regulatory elements associated with the respective gene. As a binary matrix an entry has the value 1 if the given gene has, within its upstream region, the given motif. Otherwise the entry receives the value 0. The blue dots in the representation are those entries of value 1.**

them through the program MatInspector [96], relying on the default parameters. Figure A.5 depicts a binary sparse matrix (See Table A.1) represented in Matlab [78], in which the genes in the dataset lie on the vertical axis and the potential regulatory elements on the horizontal.

Appendix B

IUPAC Ambiguity Codes

Symbol	Meaning
G	G
A	A
T	T
C	C
R	G or A
Y	T or C
M	A or C
K	G or T
S	G or C
W	A or T
H	A, C or T
B	G, T or C
V	G, C or A
D	G, A or T
N	G, A, T or C

Table B.1: **Single letter codes for ambiguous nucleotides defined by the International Union of Pure and Applied Chemistry (IUPAC) [25].**