# COMPUTATIONAL BIOLOGY APPROACHES FOR STUDYING GENE REGULATORY NETWORK DISCOVERY AND MODELLING

by

RAFIK A SALAMA

A thesis submitted to

The University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

School of Biosciences

The University of Birmingham

September 2011

# Abstract

The advent of next generation sequencing has increased the gap between genome sequence data and knowledge, enhancing the need for faster means to fill this gap. The development of efficient computational biology methods to handle this gap has never been so important. Gene regulatory networks in particular have been studied widely for their role in controlling cellular behaviour, resulting in manifold phenotypic characteristics. In this thesis, I present novel techniques contributing to the discovery of gene regulatory network connections, through enhanced binding site prediction, binding site multiple sequence alignment and binding site specificity. Another major advantage of computational biology is the ability to simulate the behaviour of gene regulatory networks, in order to study the governing dynamics of such networks. In this thesis, I also introduce a new modelling language bringing computational modelling capabilities into the biological domain to simplify the process of writing a model that can be simulated *in silico*. I have proved through this work that: first, the devised computational biology techniques can provide cheap yet powerful and efficient techniques to study gene regulatory networks; and second, the techniques presented have novel superiority over current research in this domain.

# Acknowledgements

Many thanks to my family and friends for their support during the last three years. Special and countless thanks go to my son Ghali and my wife Maggie, for going out of her way - leaving her family, friends and high profile job back in our home country Egypt to join me in this journey for my PhD. Many more thanks for her support along the way through tough times in research being the best friend at all times and for her efforts in constant revisions of this thesis.

Thanks to my supervisor Dov Stekel to whom I owe limitless debts for his continuous mentoring throughout my PhD and his support when I needed it most and his continuous encouragement for every step, guiding me through the right path to become an independent researcher.

Thanks for the Darwin Trust of Edinburgh who sponsored me fully from the time I stepped here in England till the time I handed in my PhD thesis.

Special mentions goes to Dr. Jan Kreft, Dorota, Dafyd and countless others who helped from the CSB and elsewhere in the university.

# CONTENTS

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| **ATP** | Adenosine Tri Phosphate |
| **AUC** | Area Under Curve |
| **BLOSUM** | BLOcks of Amino Acid SUbstitution Matrix |
| **CAP** | Catabolite Activator Protein |
| **ChIP** | Chromatin Immuno Precipitation |
| **CRP** | cAMP Receptor Protein |
| **DBD** | DNA Binding Domain |
| **ddNTP** | Dideoxynucleotide Tri Phosphate |
| **DNA** | Deoxyribonucleic Acid |
| **EDNA** | Thermodynamic Energy based Di Nucleotide multiple sequence Alignment |
| **EM** | Expectation Maximization |
| **FDR** | False Discovery Rate |
| **HMM** | Hidden Markov Model |
| **mRNA** | Messenger Ribonucleic Acid |
| **MSA** | Multiple Sequence Alignment |
| **NGS** | Next Generation Sequencing |
| **PAM** | Point Accepted Mutation |
| **PCR** | Polymerase Chain Reaction |
| **PSWM** | Position Specific Weight Matrix |
| **RNA** | Ribonucleic Acid |
| **RNAP** | Ribonucleic Acid Polymerase |

| | |
|---|---|
| **ROC** | Receiver Operating Characteristic |
| **SBML** | Systems Biology Mark-up Language |
| **SDNMSA** | Statistical Di Nucleotide Multiple Sequence Alignment |
| **SOLiD** | Sequencing by Oligonucleotide Ligation and Detection |
| **TFBS** | Transcription Factor Binding Site |
| **UJP** | Ungapped Joint Probability |
| **ULPB** | Ungaped Likelihood under Positional Background |
| **UPGMA** | Unweighted Pair Group Method with Arithmetic Mean |

Chapter 1

# INTRODUCTION

Jacques Monod and Francois Jacob were the first to describe a gene regulatory mechanism (Jacob and Monod, 1961): the enzymes involved in the metabolism of lactose in *Escherichia coli* are only expressed in the presence of lactose and in the absence of glucose. From then, researchers started to think of genetic connections, where the protein of one gene may regulate another gene, which was later described as a network (Britten and Davidson, 1969). Gene regulatory networks have become the focus of much research, trying to elucidate the genetic connections and control structures resulting in numerous phenotypic characteristics (Strohman, 2002).

From a network perspective, a gene network associated with a certain gene "A" includes those genes that are either activated or repressed by the protein molecule translated from gene "A". The relationship between genes in that sense can be likened to control circuits, or thought of as a cause and effect relation or a source and target relation. In general, the source regulatory genes are classified by their regulatory role. An activator is a gene that translates into a protein that increases the propensity for a target gene to be transcribed and translated; a repressor decreases the propensity of a target gene being transcribed and translated. The modes of interaction can vary from indirect to direct interactions. An example of an indirect interaction is where a repressor does not bind physically to the DNA but intervenes with an important

activator preventing it from binding to the site. In contrast, a direct interaction could be when the repressor binds to the DNA repressor site preventing RNA polymerase from initiating transcription.

By way of illustration in **prokaryotes**, lactose metabolism in *E. coli* (Jacob and Monod, 1961), lactose sugar molecules act as a indirect repressors for the direct repressor transcription factor that represses the production of β-galactosidase enzymes. If lactose is missing from the environment, the direct repressor binds to the DNA, preventing the production of this enzyme. On the other hand, the absence of glucose increases the number of cyclic AMP molecules, and indirectly activates β-galactosidase enzyme production. This is mediated by cAMP binding to Catabolite Activator Proteins (CAP), which directly bind to the DNA on the CAP binding site. This increases the propensity of RNAP binding to DNA, and hence increases the production of β-galactosidase enzyme. This interaction is an example of simple prokaryotic gene regulatory network that responds dynamically to changes in environmental food sources.

The gene regulatory interactions vary considerably between prokaryotes and eukaryotes. In **prokaryotes**, gene regulation is mainly initiated by the binding of RNA polymerase to one of the sigma factors (Sharma and Chatterji, 2010, Gruber and Gross, 2003, Kapanidis et al., 2005, Helmann and Chamberlin, 1988) to form a holoenzyme which then binds to the promoter region to initiate the transcription. This process is controlled by another protein molecule called transcription factor (Seshasayee et al., 2011, Khan and Kumar, 2009, Latchman, 1997). Transcription factors in general are protein molecules, which bind to the DNA and either facilitate or block the RNA polymerase to bind to the promoter region to start transcription. The transcription factor interaction can be direct by facilitating the recruitment of the

holoenzyme to the DNA or indirect by cooperating with other factors. Transcription factors are characterized by the presence of specific domain that assists the binding event to the DNA, termed the DNA Binding Domain (DBD) (Babu et al., 2004). The binding of this protein molecule to the DNA involves two major interactions between the DNA binding site and the transcription factor DBD. First, direct read outs which represents the direct interactions between the amino acids and the DNA bases through specific hydrogen bonds; second, indirect readouts which result between indirect interactions with the side chains of the protein enabled by conformational changes in the DNA structure. Accordingly; transcription factors have attracted a lot of attention for their importance in revealing the network of genetic regulation in multiple species. In particular, I have extensively studied *E. coli* where there are almost 170 transcription factor known as per the RegulonDB, those transcription factors have been extensively studied, of which more than 20 factors regulate more than 20 genes that I will study in this thesis in details. Of those 20 transcription factors, three were studied in details. These are the cAMP-receptor protein (CRP), LexA and ArcA.

**CRP**: cAMP-receptor protein (CRP) is one of the seven ‘‘global’’ transcription factors in *E. coli* (Martinez-Antonio and Collado-Vides, 2003). It is known to regulate more than one hundred transcription units (Jacob and Monod, 1961). CRP’s activity is triggered by binding of the second messenger cAMP in response to glucose starvation and other stresses (Jacob and Monod, 1961). CRP binding sites have proved to be particularly noisy as the computational searching for the consensus binding site can easily miss lots of known binding sites. CRP was chosen for its high promiscuity to the transcription factors.

**LexA**: LexA directly regulates ~30 *E. coli* transcription units involved in the "SOS" response (Walker, 2000). Such transcription is induced in response to DNA damage. Under normal growth conditions, LexA binds to a specific 20-base-pair (bp) sequence within the promoter regions of these genes, repressing transcription by sterically occluding RNA polymerase (RNAP). LexA was chosen for its lower promiscuity to the transcription factors, which should exhibit better behaviour than the CRP binding site.

**ArcA**: ArcA is a global regulator that changes in relation to the expression of fermentation genes and represses the aerobic pathways when *Escherichia coli* enter low oxygen growth conditions (Nikel et al., 2008). ArcA was chosen for its different protein domain (CheY like) and a very low consensus of the binding site.

**Eukaryotic** gene regulation on the other hand is much more complex than Prokaryotic gene regulation that involves more complex interactions and factors that control the transcription and translation of the genes. In Eukaryotes, the gene regulation adds extra layers of regulations from the transcription of genes to their translation if they are coding genes. Initially there is a control on the chromatin level where the genes can be either activated or silenced for transcription by the chromatin states (epi-genetics) (Russo et al., 1996, Bird, 2007, Rosenfeld et al., 2009). The compaction of chromatin into heterochromatin indicates inaccessibility of the ribosomal proteins to initiate the process (Holliday, 1990). It has been found that the N terminus of the histone proteins undergoes chemical modifications such as acetylation and methylation indicating the transcription state of the gene (Strahl and Allis, 2000). For example, if the Lysine residue 9 in the histone protein 3 is tri methylated (H3K9me3) then this is gene is activated and the code H3K27me3 indicates repression (Jenuwein and Allis, 2001, Hublitz et

al., 2009, Barski et al., 2007, Koch et al., 2007, Wang et al., 2010), etc. The second level of control is transcription factors. These operate in an analogous way to transcription factors in prokaryotes, induce or repress the transcription by either assisting/blocking the RNA polymerase to bind to the promoter region of the gene. Other distinct players in the Eukaryotes are the enhancers that are known to bind far from the gene being regulated (Spilianakis et al., 2005). Some are even found hundred of thousand base pairs from the start site and some are even not on the same chromosome of the gene (Arnosti and Kulkarni, 2005). They are also known to bind in Introns, which explain the effect of intron polymorphism on gene regulation. The enhancers are known to regulate indirectly by interacting with the transcription factor using the super coiled structure of the chromatin through spatial proximity. This interaction enhances the recruitment of the RNA polymerase.

The transcription results mainly in pre mature messenger RNA (pre-mRNA). It contains two types of sequences, exons that are translated as part of the protein being synthesized and introns, which are not translated. In this level of regulation, the micro RNA (Bartel, 2009, Bartel, 2004, Lagos-Quintana et al., 2001) a famous non-coding RNA molecule comes into play.. The Micro RNA are small molecules that bind to complementary sequences in the pre mRNA to repress its translation (He and Hannon, 2004).

The science of gene network interactions has sparked many research branches. This thesis will focus on **prokaryotes** in particular. **First,** to elucidate the gene network structure, we need to **discover the regulatory interactions** between various genes. These activities include discovering the binding sites for transcription factors either biologically or computationally. This is done by aligning those binding sites efficiently and studying the binding sites'

specificities. **Second, discovering the dynamics governing a network**, ideally through computational/mathematical modelling, simulations and analyses of such networks.

Accordingly, this thesis is organized into two parts: **Part I**, focusing on research into binding site discovery and alignment; and **Part II**, focusing on the modelling of gene regulatory networks. In addition, preliminary work on the specificity of binding sites is included into the thesis and exhibited in Appendix I. There follows two brief introductory sections for the two thesis parts, and a summary of each chapter.

## 1.1. Part I: Discovery of Gene Regulatory Networks

### 1.1.1 Introduction

The discovery of connections between genes implicitly identifies the regulatory network between them. Accordingly, the prediction of binding sites for a certain transcription factor provides a useful route for the discovery of gene regulatory networks. Hence, in this thesis I have focused on the prediction and alignment of transcription factor binding sites, introducing novel techniques that have proved to be superior to other current methods. There follows a brief review of research into both binding site prediction and multiple sequence alignment.

#### 1.1.1.1. Prediction of Binding Sites

The prediction of binding sites for a given transcription factor involves a number of techniques, both biological and computational. Biological techniques have proved to be accurate but costly in terms of resources and time needed to confirm a binding site.

##### A. Biological Techniques

**DNA foot printing** is an important, low-throughput, technique originally developed to assess the binding of one molecule to the DNA region of interest (Leblanc and Moss, 2001). This technique can be summarized as follows:

1. The DNA binding region of interest is amplified through polymerase chain reaction (PCR).

2. The amplified DNA sample is split into control and experimental sample.
3. The experimental sample is mixed with the protein of interest and left to bind.
4. Both the sample and control are cleaved using a cleavage agent, which cuts the DNA molecule in random locations. The areas where the protein is bound will not be cleaved as they are protected against cleavage by the bound protein.
5. Both the sample and control can be tested for different cleaved points of the DNA. One will observe the points cleaved in the control sample that is not cleaved in the experimental ones. This can be shown using gel electrophoresis (Berg et al., 2007), which will show an area with no bands when run on the gel.

On the other hand, high throughput techniques have been used to deliver a complete DNA binding distribution of a specific transcription factor. For example, the **ChIP-on-chip** (Aparicio et al., 2005, Ren et al., 2000) relies on the combination of microarray technology with chromatin immunoprecipitation; where:-

1. The transcription factor in question is cross-linked with the DNA molecule.
2. The DNA is then sonicated to chunks of 1kilobase pair or less.
3. An antibody specifically designed for the transcription factor is used to recover immune precipitated DNA-protein complex. This will lead to identifying where about those filtered DNA-protein complex is on the DNA.
4. The complex is reverse cross-linked and the single strand DNA is obtained, amplified and denatured.
5. The DNA strand is tagged with a fluorescent tag.

6. The final step is pouring the labelled DNA strand fragments over the complementary DNA strands of known DNA positions, arranged using a DNA array. Hybridization is then identified by measuring the fluorescence signal along the DNA. The resulting fluorescence image is analyzed computationally to identify the binding positions.

With the advancement of **next generation sequencing (NGS)** as will be explained later, the use of short tag reads substitutes for microarray technology and is used to detect the position of the binding site, resulting in a new method **ChIP-seq** (Johnson et al., 2007). The protocol of ChIP-seq can be summarized as follows:

1. The transcription factor in question is cross-linked with the DNA molecule.
2. The DNA is sonicated to chunks of 1-kilo base pair or less.
3. An antibody specifically designed for the transcription factor is used to immunoprecipitate the DNA-protein complex.
4. The immunoprecipitated DNA fragments can be amplified and directly sequenced using one of the NGS technologies.
5. The sequencing produces thousands of overlapped short DNA tags (up to 50-mer). These sequences can be then matched computationally against the template genome to identify the start and end positions of each tag and the binding site can be detected up to a high resolution.

The different NGS technologies themselves can be summarized as follows:

**Sequencing by synthesis** as used lately in pyrosequencing (Ronaghi et al., 1998). This method relies on synthesizing a complementary DNA strand to the unknown strand. It depends on pyrophosphate released from the hybridized nucleotide, which is then used to generate an ATP molecule that mediates a light emission reaction using oxyluciferin. 454 parallel pyrosequencing machines incorporate a parallel version of this reaction using emulsion PCR, which generates millions of 200-400 bases, reads.

**Sequencing by dye termination (Mardis, 2008)** which follows the same principle method of chain termination as in the Sanger method. Sanger sequencing uses a collection of differentially labelled dideoxynucleotides triphosphates (ddNTPs, which are DNA bases with 3' blocker, not allowing for extension after being bound to the DNA), and then analyze it using gel electrophoresis. A better technique uses ddNTPs modified by conjugation to fluorophores emitting light of a different wavelength for each base. The result is a generation of different fluorescence signals for each ddNTP, which can then be analyzed to pinpoint the position of each nucleotide.

Illumina (Bentley et al., 2008) uses a more advanced version of the later technique. The dye fluorescence is reversible, and the terminal 3' end blocker is removed after each addition, so that 4 ddNTP types (A,C,G,T) are added each at a time after amplifying the DNA tag to be sequenced then an image is taken to identify the ddNTP. Later, the terminal 3' blocker along with the dye is removed, allowing for another cycle of addition until the DNA is sequenced.

**Sequencing by ligation (McKernan et al., 2009)** relies on the DNA ligase to add a library of fluorescent oligonucleotides to the unknown DNA strand. This method relies heavily on the

efficiency of DNA ligase to detect DNA mismatches. SOLiD uses a parallel amplified (using emulsion PCR) version of this protocol. SOLiD is known to read up to 2-4 billion base per run.

## B. Computational Techniques

Although biological techniques have proved to be efficient in discovering binding sites, they have also proved to be lengthy and costly. For example, using a 454 sequencing machine would cost approximately £1000 to sequence a bacterial genome (4.6 Mbase) with a 10 run coverage , and would take up to a month, despite manufacturers' claims that it should take 8 days (454-Sequencing, 2011). Using Illumina technology can be slightly cheaper and faster, but will still be in range of a week and hundreds of pounds.

On the other hand, these recent advances in sequencing technologies have resulted in the creation of a vast amount of genomes ready to be analyzed and studied. This is illustrated in Figure 1-1 with data from the GOLD database, showing the increase in the number of sequenced genomes and projects in progress.

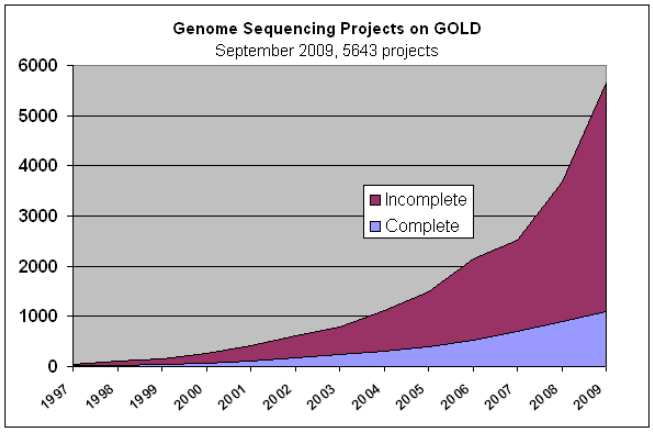


**Figure 1 - 1**: Genome sequencing projects on GOLD showing the completely sequenced genomes in blue and the incompletely sequenced genomes in red up until last year. (Figure is taken from http://www.genomesonline.org/images/gold_s2.gif)

Accordingly, there is a clear and important need to provide rapid ways of discovering and analyzing gene regulatory networks from newly sequenced genomes. This need can be addressed in part by computational approaches, which can include better prediction methods that can help with biological annotations of the data. Computational methods for prediction of TFBSs fall into two broad classes:

i. **De novo binding site prediction,** in which upstream regions of genes believed to be co-regulated are analyzed for over-represented motifs and thus typically identifying binding site motifs without using prior knowledge of known binding sites (Tompa et al., 2005). Those co-regulated genes can be mainly discovered by using various means, either computationally searching for the genes that share a functional category or biologically using ChiP-on-chip or ChiP-seq or any other genome wide discovery method. The promoter regions (or the ChiP peak sequences) of those genes are then selected for finding over represented motifs. Multiple methods have been developed over the past years to discover those over represented motifs. These methods can be classified as follows :

   a) Positional bias, using the concentration of a motif near the transcriptional start site (Hughes et al., 2000), such as:

      I. MEME (Bailey and Elkan, 1994) that assumes the presence of multiple sub-motifs in the dataset and scans the sequences for windows of size W looking for motifs distributed by a mixture of multiple sub population distributions. MEME fits a mixture model to the various motifs of binding sites using an Expectation Maximization (EM) approach.

II. On the other hand, some of the Gibbs Sampler algorithms (Thompson et al., 2007, Thompson et al., 2003, Lawrence et al., 1993, Lawrence and Reilly, 1990, Liu and Lawrence, 1999, Martin and Orkin, 1990) use a Bayesian approach to sample from the unknown motif distribution using Gibbs sampling technique for local multiple of sequences to identify similar patterns or motifs in the promoter regions; Weeder (Pavesi et al., 2001), AGLAM (Kim et al., 2008). Other tools rely on aligned sequences and use phylogeny to distinguish between motifs resulting from evolutionary proximity and motifs representing functional sites as PhyloGibbs (Storms et al., 2010, Siddharthan, 2008, Siddharthan and van Nimwegen, 2007, Siddharthan, 2006, Siddharthan et al., 2005).

b) Group Specificity: most of the tools discovering motifs end up discovering over represented motifs in the promoter region that are over represented in most of the genome. Hence, a set of tools have been developed to compare the localization of motifs in coding regions rather than non coding regions (Hughes et al., 2000), like DME(Smith et al., 2005), DEME (Redhead and Bailey, 2007), Seeder (Fauteux et al., 2008) using discriminative analysis, or using Least likelihood under background model (Friberg et al., 2005).

ii. **Training based methods** that rely on a training set of known binding sites to detect other binding sites. These methods typically rely on scoring positions of a training set either statistically or energetically. Training based methods can be classified as:

a) Consensus based methods using the position weight matrix (Hertz and Stormo, 1999). These methods mostly use the position specific weight matrix (PSWM) that describes the frequency of base occurrence (A, C, G, and T) in each position of an alignment. PSWM is computed as $P_i(x)$ for [A, C, G, T] at each position $\boldsymbol{i}$ from $f_i(x)$, the frequency of each base $\boldsymbol{x}$ among the sequences (that may include a pseudo-count to compensate for under sampling (Durbin et al., 1998)). Accordingly, if there are N sequences in the alignment (with appropriate pseudo-count correction), the proportion of symbol $\boldsymbol{x}$ in position $\boldsymbol{i}$ is given by $P_i(x) = f_i(x)$, hence, given a new sequence of symbols$(x_i, \ldots . x_m)$, the simplest measure of position specific probability associated with this sequence is: $\prod_{i=1}^{m} P_i(x_i)$

b) Bayesian modelling of the binding site positions (Merkulova et al., 2007, Osada et al., 2004), where the models represent the binding sites position using Bayesian networks indentifying the inter dependencies between the binding site positions. This is shown in the method devised by Ben-Gal, where he used a Variable Order Bayesian Networks (Ben-Gal et al., 2005) to represent multiple orders interdependencies between positions. Biophysical methods, as QPMEME (Djordjevic et al., 2003), using the binding energies between the amino acids and the DNA bases to calculate the likelihood of a binding site.

c) Hidden Markov Models (HMM) of binding site positions: the normal PSWM can be also called the ungapped score matrix as it does not allow for evolutionary insertions or deletions represented by gaps in a multiple sequence alignment (MSA) into the computation of the score. The score will typically be calculated for all appropriate sub

sequences of an upstream region in order to identify the most likely binding sites. As mentioned previously, incorporating gaps into MSAs to allow representation of insertions or deletions has been found to increase the specificity of alignment models (Durbin et al., 1998). Therefore, an evolutionary derived gapped model of the training sequences might provide a better prediction of the binding site likelihood. One way to achieve a gapped model of the binding site is with a Hidden Markov Model (HMM) (Durbin et al., 1998). HMMs have been used previously in research of binding site prediction to assess the likelihood of the binding site based on its statistical evolutionary profile.

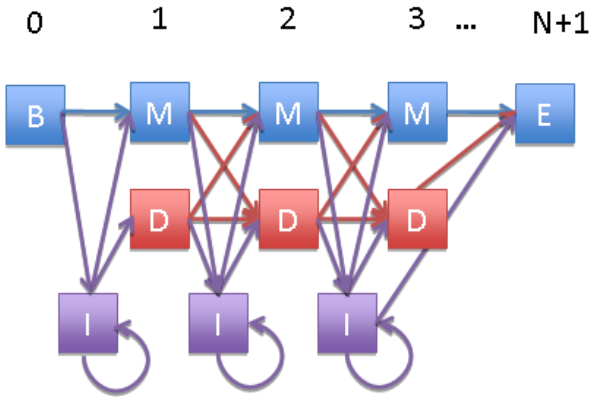


**Figure 2 - 2**: A zero order Hidden Markov Model (HMM) with only 3 states, first state "M" for matching a nucleotide base in the observed sequence, "I" an insertion state for inserting a nucleotide base between match states and "D" a delete state representing no nucleotides observations. Two special states termed "B" as the begin state and "E" to represent the end state.

A zero order HMM models the sequence of bases as a Markov chain of mainly 3 states (Match, Delete and Insert) as described by Durbin et al. (Durbin et al., 1998). Transition and emission probabilities are calculated using an MSA of the training set of sequences.

#### 1.1.1.2. Multiple Sequence Alignment

Multiple sequence alignment is a useful tool for researchers. Originally developed to identify similar regions between sequences, where it is an important preliminary step to computational prediction of transcription factor binding sites, as an alignment is needed to identify the correct consensus, which will be used as an input to any prediction algorithm. The majority of research on multiple sequence alignments has been directed towards aligning protein sequences, although the same approaches have been also used to align DNA coding regions.

The process of alignment involves aligning equivalent residues to achieve homology relationship between sequences. The simple alignment of two sequences (pairwise alignments) can be carried using an iterative approach relying on Levenshtein or edit distance (Levenshtein, 1966): this allows both for mismatches and gaps, with the final score representing the distance between the two sequences. A more robust approach can be carried out relatively quickly, using efficient algorithms. Two main algorithms have been devised to obtain an optimal pair-wise alignment using dynamic programming in O(nm): Smith-Waterman for local alignment (Smith and Waterman, 1981) identifying local residues alignment and Needleman-Wunsch for global alignment (Needleman and Wunsch, 1970) aligning every possible segment of sequences and choosing the optimal arrangement. A hybrid approach, or semi-global alignment, has also been described (Brudno et al., 2003c), which considers a global alignment to edit the sequences while showing the local overlapping regions. For searching large databases for aligning sequences, faster approaches to pair-wise alignments rely on heuristics, for example BLAST that uses word shift algorithms (Altschul et al., 1990).

Aligning more than two sequences (Multiple Sequence Alignment) is a computationally intensive task as the complexity of the algorithm increases exponentially with the number of sequences involved with an NP-hard complexity (Just, 2001, Wang and Jiang, 1994). Accordingly, most MSA programs rely on heuristic approaches to speed up the process of alignment. Typically, the multiple sequence alignment follows four steps as shown in figure 1-3

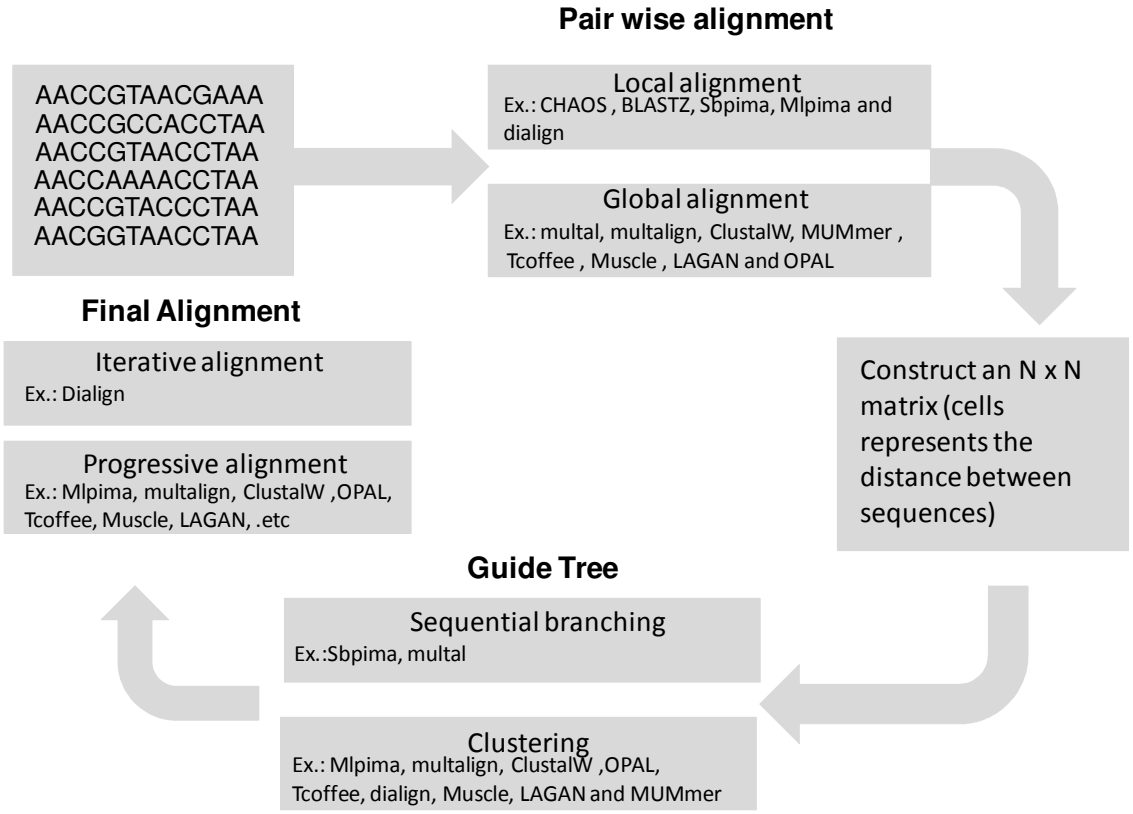


**Figure 1 - 3**: A figure representing the various stages for multiple sequence alignment annotated by the current tools including these stages

The alignment starts normally by pairwise alignment of every tuple either using local alignments as CHAOS (Brudno et al., 2003a), SBpima, MLpima and BLASTZ (Schwartz et al.,

2000) or global alignments as AVID (Bray et al., 2003), MUMmer (Delcher et al., 1999),ClustalW (Thompson et al., 2002), TCoffee (Notredame et al., 2000), Dialign(Morgenstern, 2007), MAFFT (Katoh et al., 2005), Muscle (Edgar, 2004), LAGAN (Brudno et al., 2003b) and OPAL (Wheeler and Kececioglu, 2007). The next typical step is to calculate the distance between the aligned pairs as the one used in ClustalW for instance.

$$D = 1 - \frac{N_{identical}}{N_{aligned}}$$

Where D is the distance between the aligned sequences and $N_{identical}$ is the number of identical residues and $N_{aligned}$ is the number of aligned residues.

Following the construction of the distance matrix, the final alignment can be constructed either progressively where sequences are added one by one or iteratively where sequences are iteratively split into n groups then aligning those groups for m times. The progressive alignment relies on the evolutionary relationship between the sequences to determine the right order for aligned sequences. The evolutionary relationship between the sequences uses the distance matrix to cluster the sequences based on their proximity through a guide tree. This tree can be generated either through a sequential branching as in SBpima (Smith, 1992) or through clustering using either a Neighbor joining algorithm (Zhang and Sun, 2008, Saitou and Nei, 1987) or Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Sokal R, 1958) .

The alignment algorithm in general requires other inputs as the substitution matrix and gap penalties. Considerable research has been carried out into suitable **substitution matrix** that determines the probability of transformation from one amino acid to another, or insertion or deletion mutations. Dayhoff (Dayhoff, 1978) developed one of the first substitution matrices

looking at 71 families of closely related proteins and considering amino acid accepted mutation (Point Accepted Mutation, PAM) then estimated the substitution matrix for a given evolutionary interval. Another important matrix was then developed by Henikoff (Henikoff and Henikoff, 1992) in 1992 which have calculated the substitution matrix based on divergent protein families (rather than related ones) by looking at the conserved regions in the protein BLOCKS database (Henikoff and Henikoff, 1991) and calculating the odds of the amino acid substitution in homologous sequences. Other advanced matrices have been developed and which consider structural information or a specific context (Teodorescu et al., 2004, Tyagi et al., 2006). DNA substitution matrices on the other hand have undergone a different direction of research. DNA bases belong to two groups, Purine and Pyrimidine; therefore, two substitution rates were defined. These are, transition, which is a substitution between two purine rings (A – G) or two pyrimidine rings (C- T), and transversion, which is a substitution of a purine with pyrimidine and vice versa (Brown et al., 1982, Gojobori et al., 1982, Curtis and Clegg, 1984, Wakeley, 1994, Wakeley, 1993). These rates have been calculated using various training sets of DNA sequences (Purvis and Bromham, 1997, Yang and Yoder, 1999, Lanave et al., 1986, Ina, 1998, Strandberg and Salter, 2004).

**Gap Penalties** represent another important factor of an alignment algorithm, since they control the penalty to open a gap within the sequence (internal gaps) or at the start and end of the sequence (external gaps). There are several models for gap penalties, the simplest of which is a fixed penalty. Another model is a linear model representing the gap penalty per unit length of gaps. The most commonly used approach is the affine gap penalty that represents the gap as a function of two parameters, open gap penalty and extension gap penalty (Altschul and

Erickson, 1986). The significance of the chosen values for these gap penalties has attracted a lot of research for better optimization techniques (Vingron and Waterman, 1994).

In addition to alignment methodologies, researchers require tools to edit the alignment manually to correct any unfavoured automatic alignment and to visualize the alignment through semantically representative colours. These and other features have been incorporated into Jalview (Waterhouse et al., 2009).

### 1.1.2 Part Structure

**Part I** focuses on three research foci that are split into three chapters; binding site prediction (Chapter 2); binding site alignment (Chapter 3) and enhancing the prediction power of ChIP-on-chip methods (Chapter 4).

All of the material presented in Chapter 2 and some of the material presented in Chapter 4 have been adapted from published work: Salama, R.A. and Stekel, D.J. 2010, Inclusion of neighbouring base interdependencies substantially improves genome-wide prokaryotic transcription factor binding site prediction. *Nucleic Acids Research*, 38**: e135.** I wrote the text of the paper and developed the software, while Dov Stekel was responsible for reviewing and editing the paper before publication.

#### 1.1.2.1. Chapter 2

In this chapter, I have considered two models for enhancing binding site prediction. Both models introduce a novel approach that considers interdependencies between neighbouring DNA bases: first, through an ungapped alignment model of the binding sites; and second, considering a gapped model which relies on multiple sequence alignment as preliminary step.

**A. Incorporating Binding Site Base Dependencies in <u>ungapped</u> Prediction algorithms.**

Most of the published methods for binding site prediction rely on position specific weights that assume independence between binding site bases. However, it has long been known that the interactions between neighbouring DNA bases have a significant impact on DNA topology. A major hypothesis behind such dependence is the thermodynamic properties of base stacking interactions which have been extensively measured, and are commonly used in computational methods for DNA secondary structure prediction (Mathews et al., 1999) . Another idea behind this hypothesis is that of compensating mutations between neighbouring DNA bases (Stormo et al., 1986). Tomovic and Oakeley have also shown that there are statistical dependencies between bases and that they correlate with DNA structure (Tomovic and Oakeley, 2007). Accordingly, I have modified the independent prediction models and devised a set of new models modified to incorporate the dependency in binding sites predictions as shown in chapter (2).

## B. Incorporating Binding Site Base Dependencies in gapped Prediction algorithms.

Multiple Sequence Alignment (MSA) has been long used to align sequences; these are commonly produced using methods that assume evolutionary relatedness between the sequences being aligned. They typically work by using gaps to represent insertions or deletions and substitution matrices to represent point mutations, as will be explained in detail in chapter 3. These gaps have been found to increase the specificity of alignment models (Durbin et al., 1998). Therefore, a gapped model of the training sequences might provide a better prediction of the binding site likelihood. One way to achieve a gapped model based prediction of binding

sites is with a Hidden Markov Model (HMM) (Durbin et al., 1998). Since most published papers assume independence of DNA bases, I have devised a new first order Hidden Markov Model incorporating base dependencies and testing its performance against ungapped prediction algorithms. This has shown to be inefficient compared with the ungapped model, primarily because of its high sensitivity to the training MSA used, and the fact that all the MSAs themselves assume independence between binding site positions. This result has inspired the work of Chapter 3.

#### 1.1.2.2. Chapter 3:

In this chapter, I describe implementations of enhancements to the multiple sequence alignment programs, by introducing a new approach for aligning binding sites that are not necessarily evolutionarily related. I present two novel approaches for the alignment of transcription factor binding sites, the first of which is based on interdependence between the binding site bases and the second considers the stacking free energy of the interdependent bases. Both models are tested using their prediction power for binding sites, making use of the gapped first order HMM described in Chapter 2.

**A. An interdependent statistical approach to alignment of transcription factor binding sites**

MSA tools currently used use substitution matrices and gap penalties derived based on evolutionary relatedness between the sequences aligned. This assumption does not necessarily apply for aligning a set of binding sites targeted by the same transcription factor. This

deficiency in MSA method has led us to devise a new MSA approach that incorporates neighbouring dependencies between DNA bases across the binding site positions. Since the binding sites do not necessarily have the same DNA sequence (as in most cases), I hypothesize that some of the variation among the binding sites could be selected to conserve DNA base interdependence. Hence, I present a new model that encodes the DNA binding sites into a di-nucleotide character base and calculate the substitution matrix using an approach analogous to expectation maximization.

### B. A Thermodynamic Approach To Alignment of Non-Coding regions (Transcription Factor Binding Sites).

Interdependence between DNA bases could be explained by base stacking interactions between the bases. Hence, I present a second approach that considers stacking free energies between DNA bases as a different null hypothesis from the one used in the previous model. I will show that this model is superior in its prediction power relative to the first model for most of the binding sites considered, while providing same behaviour for highly conserved binding sites.

## 1.1.2.3. Chapter 4:

The final chapter of Part I considers the enhancement of binding site prediction. In this chapter, I have investigated the ability to harness the prediction power of both ChIP-on-chip and computational prediction in an attempt to enhance the power of prediction of both approaches, and provided a collective model indicating linear regression between ChIP-on-chip signal and likelihood functions.

## A. Combining Prediction Likelihood with ChIP-on-chip signal to improve accuracy.

ChIP-on-chip signal is known for its low precision in identifying the binding sites, as it is limited by target size and probe density, so that a binding event may only be detected in a range of 1000 base pairs, which can be hard to narrow down statistically. Accordingly, combining ChIP-on-chip signals with prediction likelihood of binding sites could significantly narrow down the detected subset of binding sites by either approach alone. Hence, I have devised a method to combine ChIP-on-chip signals with computational prediction likelihood to provide better criteria for narrowing down the predicted binding sites.

## B. A collective model for linear regression between ChIP-on-chip signal and statistical likelihood

A linear regression between the ChIP-on-chip signal and the prediction likelihood has not been successful due to the low resolution of the ChIP-on-chip signal. To obtain a meaningful correlation between the signal and likelihood, I have used a collective model where the collective signal (sum of signals) in a range DNA base pair positions is regressed against the collective likelihood of the same range. This approach has overcome the low resolution of the ChIP-on-chip signal and indicated an apparent linear correlation.

## 1.2. Part II: Modelling Gene Regulatory Network

### 1.2.1 Introduction

Studying the dynamics of gene regulatory networks allows the researcher to understand the behaviours of the network under different conditions. Such a study can be conducted by various methods, including deterministic and stochastic models.

Model analysis typically involves either solving a particular cellular adaptation problem under various conditions, or understanding the importance of system components. Conditions for adaptation problems can be either an environmental stimulus, such as for example, modelling heat shock response, as in the work of (Srivastava et al., 2001) modelling the σ32 stress circuit in *E. coli*. using Stochastic Petri Nets, (Goss and Peccoud, 1998), the work of (Swinnen et al., 2006), or modelling the metal metabolism (Curis et al., 2009), or modelling a change in the medium, for example switching between normal medium to acidic medium as in the work of (Ross et al., 2003) and (Presser et al., 1998) modelling the effect of pH on the growth of *E. coli*, or modelling quorum sensing as in the work of (Li et al., 2006).

All these conditions can be factored in a model; however, the responding genes or pathways have to be represented as scenarios that respond to the changes. For example, in the case of *E. coli*, modelling the Lac operon along with changing the sugar type from glucose to fructose in SBML would require representing the gene regulatory network as a set of chemical reactions. Those reactions represents initially the reaction of Lac repressor repression on lactose present and the CAP activation in absence of glucose as in the toolbox developed by (Becker et al.,

2007) in the COBRA toolbox in matlab. What a modeller would typically aim from such a model is obtaining at first a baseline from his/her model that fits the actual set of data obtained in experimental results, and then to look at scenarios representing modified conditions. For instance, introducing both the types of sugar at the same time, and then examining the gene expressions levels that would result from such scenario. The modeller can also test the impact of mutations and adaptation scenarios, by making changes to the model. For example, in the work of (Atlas et al., 2008), to assess the importance of each gene to maintain the switch in food source as incorporating the genomic information in the model inspecting DNA replication. The result would be a "what if" scenario which can be very important to explain the dynamical role of each gene/protein/any other molecule being assessed. Modelling can also be used to explain the reason the networks are optimized in a particular arrangement from an evolutionary adaptation point of view, but then the model would have to factor cell survival/growth into consideration and energy consumption and so many other factors (Hua et al., 2006, Fong et al., 2005, Fong et al., 2003, Ibarra et al., 2002). However, generally in evolutionary modelling, the factors considered are highly biased towards the hypothesis that the modeller is testing.

In order to model a gene regulatory network, it is essential to provide a computational/ mathematical description of the biological processes, which translates the biological logic into a computational or mathematical logic. This could then be simulated stochastically (Priami et al., 2001, Gillespie, 1976, Calder et al., 2006, Goss and Peccoud, 1998, Li et al., 2006, Srivastava et al., 2001) or solved deterministically for equilibrium steady states (Machne et al., 2006, Lopes et al., Martinez et al., 1999, Willemoes et al., 2000). Such a model would require a formally defined modelling language which transforms the biological information into

mathematical/computational information, converting a biological scenario into a computable/solvable scenario as described by Gheorghe and Mitrana and the work of Sedwards and Mazza in Cyto-Sim (Errampalli et al., 2004, Gheorghe and Mitrana, 2004, Sedwards and Mazza, 2007).

Research into modelling languages as taken a number of routes, some of which are presented as readymade tools for modelling and simulation, and some are purely descriptive languages, such as SBML (Finney and Hucka, 2003), which presents a standard XML language to model any biochemical network. For stochastic simulations, a process calculus used originally to model mobile communication has proved to provide a better approximation for biological models and specifically biological signalling pathways and gene regulatory networks (Sangiorgi and Walker, 2001), albeit being very complex in syntax. Modelling languages and tools are still under heavy development in the research community, in order to provide better languages that can be used easily by biologists with little computer modelling knowledge, thus making them accessible to a wider set of researchers, rather than being restricted to people with strong computational and/or mathematical training. Such an aim has attracted many computer scientists to present various approaches for modelling languages (Errampalli et al., 2004, Finney and Hucka, 2003, Gheorghe and Mitrana, 2004, Guerriero et al., 2007, Sauro, 2006, Sedwards and Mazza, 2007). However, I have found that most of the modelling languages lack many of the useful features of programming languages, including generalization, reusability and encapsulation, and suffer from exponential growth of reactions as a factor of reactants. These deficiencies are summarized in detail in the introduction to Chapter 5.

### 1.2.2 Part structure

**Part II** describes work on modelling the dynamics of the networks, specifically describing a new modelling language that incorporates various missing features in current languages.

#### 1.2.2.1. Chapter 5:

**Compound oriented modelling language**

In this chapter, I present a new narrative modelling language that incorporates object oriented programming features and present it in a compound oriented modelling language. This language benefits from being a narrative style biologically intuitive modelling language. It addresses main issues in current modelling languages, including reusability, combinatorial expansions, and extensions. The language allows biologists to model simple situations, while still being capable of modelling complex relations between reactants in a highly organized reusable approach.

# PART I

# DISCOVERY OF GENE REGULATORY NETWORKS

This part presents the work on discovery of gene regulatory networks in three chapters. Chapter 2 for prediction of binding sites, Chapter 3 for alignment of binding sites and Chapter 4 considers ChIP-on-chip signals to enhance predictions and its linear regression with likelihoods.

# Chapter 2

# INCLUSION OF NEIGHBOURING BASE INTER-DEPENDENCIES SUBSTANTIALLY IMPROVES GENOME-WIDE PROKARYOTIC TRANSCRIPTION FACTOR BINDING SITE PREDICTION.

**Adapted from Salama and Stekel 2010, as described in the introduction.**

## 2.1. Abstract

Prediction of transcription factor binding sites is an important challenge in genome analysis. The advent of next generation genome sequencing technologies makes the development of effective computational approaches particularly imperative. I have developed a novel training-based methodology intended for prokaryotic transcription factor binding site prediction. Our methodology extends existing models by taking into account base inter-dependencies between neighbouring positions using conditional probabilities and includes genomic background

weighting. This has been tested against other existing and novel methodologies including position-specific weight matrices, first order Hidden Markov Models and joint probability models. I have also tested the use of gapped and ungapped alignments and the inclusion or exclusion of background weighting. I show that our best method enhances binding site prediction for all of the 22 *Escherichia coli* transcription factors with at least 20 known binding sites, with many showing substantial improvements. I have highlighted the advantage of using block alignments of binding sites over gapped alignments to capture neighbouring position inter-dependencies. I have also shown that combining these methods with ChIP-on-chip data has the potential to further improve binding site prediction. Finally, I have developed the ULPB platform: a user-friendly website that gives access to the prediction method devised in this work.

## 2.2. Introduction

Gene transcription is often controlled by transcription factors that bind to specific DNA binding sites; these either promote (activate) or repress (inhibit) the binding of RNA polymerase. To fully understand a gene's functions, it is helpful to understand the regulatory network context in which the gene participates, and that includes identifying the transcription factors that regulate it. Transcription factor binding sites (TFBSs) can be determined experimentally, e.g. using DNA foot printing (Leblanc and Moss, 2001), or using high throughput techniques such as ChIP-on-chip (Aparicio et al., 2005) or ChIP-seq (Johnson et al., 2007). However, with increased potential for high throughput genome sequencing (Hall, 2007), the availability of accurate computational methods for TFBS prediction have never been so important.

Although current state-of-the-art TFBS prediction algorithms use position-specific methods as explained in the introduction, it has long been known that interactions between neighbouring DNA bases have a significant impact on DNA topology. For example, the thermodynamic properties of base stacking interactions have been extensively measured, and are commonly used in computational methods for DNA secondary structure prediction(Mathews et al., 1999). This was illustrated in work discussing the effect of DNA flexure on the binding site affinity (Calladine and Drew, 1986). Compensating mutations between neighbouring DNA bases has been long known (Stormo et al., 1986) and Tomovic and Oakeley have also shown that there are statistical dependences between bases and that they correlate with DNA structure (Tomovic and Oakeley, 2007). I have also shown using mutual information analysis that there are

dependencies between neighbouring and distant positions of the TFBSs that I studied (see results).

Similar ideas have been applied to analyze the splicing signals in eukaryotes (Zhang and Marr, 1993, Agarwal P., 1998, Zhou and Liu, 2004). Other work includes the development of methodologies that can capture inter-position correlations using a set of training sequences (Agarwal P., 1998, King and Roth, 2003, Barash Y, 2003) and apply these correlations to de novo TFBS searches (Zhou and Liu, 2004). Bulyk et al. also showed that these correlations have an effect on the affinity of binding sites (Bulyk et al., 2002). Tomovic and Oakeley (Tomovic and Oakeley, 2007) have also introduced a statistical evaluator for the interdependence in binding site nucleotides based on ungapped joint probability distributions.

The aim of this work is to improve the computational prediction of transcription factor binding sites by developing new training-based methods that incorporate base interdependencies in effective ways. I assess their performance by comparing them with position specific approaches that are the most commonly used methods, hidden Markov models and the joint probability model of Tomavic and Oakeley.

## 2.3. Materials And Methods

### 2.3.1 A Novel Method for TFBS Prediction using base pair dependencies

The core of this work is the development and evaluation of three novel TFBS prediction methods that extends position-specific methods by including information about correlated changes in neighbouring DNA positions. The first method is an ungapped position specific method that makes use of a block alignment without gaps (henceforth referred to as "ungapped" methods). The second method is first order Hidden Markov Model (HMM) that is a modification of the zero order HMMs that use gapped multiple sequence alignments to account for dependencies between neighbouring positions in the binding sites. The third method is an enhancement to the ungapped model above by taking into consideration the positional background probability.

#### 2.3.1.1. Ungapped Likelihood

The ungapped scoring in this case is different from the normal position specific scoring in the sense that the probability of a base in a certain position is conditional on the occurrence of the base in the previous position. That means, the probability of finding base $x_{i+1}$ in position $i+1$ given $x_i$ in position $i$ is $\beta(x_{i+1}|x_i)$ which is computed as the frequency $f(x_i, x_{i+1})$ of finding the couple $x_i x_{i+1}$ at positions $i$ and $i+1$, divided by the frequency$f(x_i)$ of finding $x_i$ in position $i$ , so that the conditional probability of finding a base $x_{i+1}$ given the base $x_i$ is given by:

$$\beta(x_{i+1} \mid x_i) = (f(x_i, x_{i+1}) + U) / f(x_i) + 4U) \quad \textbf{(2-1)}$$

Where $U$ is a smoothing parameter that can also be thought of as a pseudo-count to compensate for under sampling (Durbin et al., 1998). I have set U = 0.25/$n$, where $n$ is the length of the alignment. The resulting matrix will contain $\beta(x_{i+1}|x_i)$ for all 16 combinations of the four bases at every position. Using this model, I am able to calculate the conditional probabilities based on a training set of known binding sites and then use these probabilities to predict the binding sites in a new sequence. Given a new sequence, the binding site likelihood is then a simple calculation of the probabilities computed over the binding site positions:

$$L_{\text{Ungapped}}(S) = p_1(x_1) \prod_{i=1}^{n-1} \beta(x_{i+1} \mid x_i) \quad \textbf{(2-2)}$$

Where$L(S)$ is the likelihood of the sequence $S$ of $n$ bases, $P_1(x_1)$ is the PSWM probability of base $x$ in position 1 and $x$ is one of the DNA bases [A,C,G,T] ).

### 2.3.1.2. Ungapped Likelihood under Positional Background (ULPB)

The second model described in our work is an enhancement over the ungapped model that considers the background sequences. In this model the background ungapped conditional probabilities for the genome of interest (e.g. *E. coli* K12 MG1655) is calculated using the entire genomic sequence so that:

$$\eta(y \mid x) = (g(x, y) + U) / (g(x) + 4U) \quad \text{(2-3)}$$

Where *y* and *x* are nucleotides [A, C, G, T], g(x) is the frequency of nucleotide x in the search sequence, and g(x,y) is the frequency of nucleotides x and y at neighbouring positions in the search sequence.

The binding sites likelihood ratio is now given as the ratio of the likelihood under the training sequence probabilities relative to the likelihood under the background model so that it becomes

$$\varphi(x_{i+1}, x_i) = \beta(x_{i+1} \mid x_i) / \eta(x_{i+1} \mid x_i) \qquad \textbf{(2-4)}$$

In addition, the likelihood is given by:

$$L_{\mathrm{ULPB}}(S) = p_1(x_1) \prod_{i=1}^{n-1} \varphi(x_{i+1} \mid x_i) \qquad \textbf{(2-5)}$$

Throughout the work, I have used the log likelihood ratios for ease of calculation.

### 2.3.1.3. First Order Gapped Hidden Markov Model

The Hidden Markov Model (HMM) used in this work is a first order HMM in which every match state emits only one base (i.e. probability of 1) and the transition probabilities capture all inter-dependencies between the binding site bases. The HMM has the usual insert and delete states associated with Profile HMMs (**Figure 2 - 1**). The HMM transition and emission probabilities are calculated using training sequences with pseudo-counts and the Viterbi algorithm.

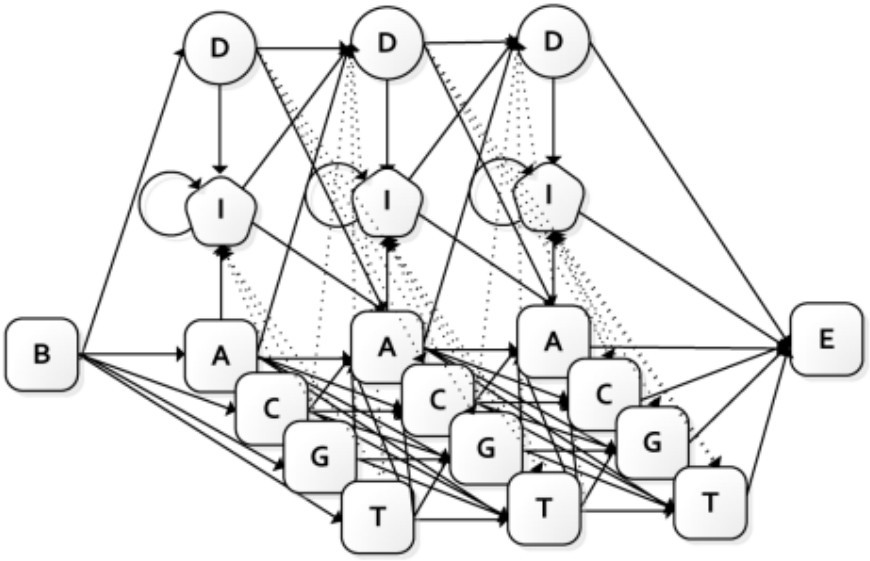


**Figure 2 - 1:** First order HMM states for the DNA sequence, with four match states [A, C, G, T] emitting A, C, G or T respectively with probability 1. D is the delete state/silent state emitting no bases and I is the insert state which emits either A, C, G or T with equal probability. B and E denotes the beginning and end states of the HMM.

The hidden markov state sequence for a given observation can be best found by finding the most probable path of states for a given observation, as defined by the Viterbi algorithm. Formally, the most probable path $\pi$ can be found recursively. If I suppose that the probability $v_k(i)$ of the most probable path ending in state k with observation I is known for all states k, then such probabilities can be calculated for observation $x_{i+1}$ as

$$v_l(i+1) = e_l(x_{i+1}) \max_k (v_k(i) a_{kl}) \qquad (2\text{-}6)$$

Where: $e_l(X_{i+1})$ The emission probability at state L for observation $i+1$, $a_{kl}$: the transition probability between state $k$ and state $L$

The Viterbi algorithm uses a dynamic programming approach to solve this problem, using the optimal substructure solution as the partial state sequence as a part of the observation. Applying the above general Viterbi equation to our First Order Markov Model, then the resulting set of equations for the states in our model is as follows:

$$V_j^A(i) = 1 + \max \begin{bmatrix} V_{j-1}^C(i-1) + \log a_{C_{j-1}A_j} \\ V_{j-1}^G(i-1) + \log a_{G_{j-1}A_j} \\ V_{j-1}^T(i-1) + \log a_{T_{j-1}A_j} \\ V_{j-1}^A(i-1) + \log a_{A_{j-1}A_j} \\ V_{j-1}^I(i-1) + \log a_{I_{j-1}A_j} \\ V_{j-1}^D(i-1) + \log a_{D_{j-1}A_j} \end{bmatrix}$$ (Same for the other three states C, G, T),

$$V_j^I(i) = \log \frac{e_{I_j}(x_i)}{q_{x_i}} + \max \begin{bmatrix} V_j^C(i-1) + \log a_{C_jI_j} \\ V_j^G(i-1) + \log a_{G_jI_j} \\ V_j^T(i-1) + \log a_{T_jI_j} \\ V_j^A(i-1) + \log a_{A_jI_j} \\ V_j^I(i-1) + \log a_{I_jI_j} \\ V_j^D(i-1) + \log a_{D_jI_j} \end{bmatrix}, \quad V_j^D(i) = \max \begin{bmatrix} V_{j-1}^C(i-1) + \log a_{C_{j-1}I_j} \\ V_{j-1}^G(i-1) + \log a_{G_{j-1}I_j} \\ V_{j-1}^T(i-1) + \log a_{T_{j-1}I_j} \\ V_{j-1}^A(i-1) + \log a_{A_{j-1}I_j} \\ V_{j-1}^I(i-1) + \log a_{I_{-1j}I_j} \\ V_{j-1}^D(i-1) + \log a_{I_{-1j}D_j} \end{bmatrix}$$

**(2-7)**

Where $V_j^Y(i)$ is the log-odds score of the best path matching subsequence $x_{1\ldots i}$ to the sub-model up to state *j,* ending with $x_i$ being emitted by state $Y_i$, where *Y* in our model can be either A, C, G, T or I. On the other hand $V_j^D(i)$ is the log-odds score of the best path ending in a silence state *D.*

### 2.3.1.4. Ungapped Joint probability (UJP)

Tomavic and Oakeley (Tomovic and Oakeley, 2007) introduced a correction to the PSWM using the ungapped joint probability of the dependant bases divided by the background probability of the bases. Assessment of their method has been made using in implementation of the scoring function shown in their Equation 22.

### 2.3.2 Assessing inter-dependency in the binding site

I have measured the inter-dependency between the binding site positions using the mutual information (Chouinard et al., 1996) between each binding site position based on the Shannon entropies at each position given by the equation

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad \textbf{(2-8)}$$

Where $I(X;Y)$ is the mutual information between position X and position Y and $H(X)$ is the entropy at position X and $H(X,Y)$ is the joint entropy between position X and position Y.

### 2.3.3 Evaluation of prediction methodologies

In our work, I have compared the training based TFBS prediction in prokaryotic binding sites between ungapped likelihood, ungapped joint probability method, ungapped likelihood under positional background, first order Hidden Markov Model (HMM) and PSWM.

For each method and each transcription factor, a leave-one-out cross validation method has been used to obtain a score for each binding sites in the training set: a model is built using all of the other binding sites for that transcription factor and that model is used to obtain a score for

the binding site in question. The *p-values* of the binding sites were calculated by comparing the leave-one-out scores with the distribution of model scores obtained for the genome sequence of *Escherichia coli K12 MG1655* using a full training-set model. The calculated *p-values* were then corrected for False Discovery Rate (Hochberg, 1995) and used to draw the ROC curves.
The ungapped (block) sequence alignments are used as suggested by RegulonDB (Gama-Castro et al., 2008) with no gaps in the sequences. I choose the orientation of the binding sites to be cis with the regulated genes, as given in RegulonDB. Multiple sequence alignment for the first order markov model was carried out using clustalw (http://www.ebi.ac.uk/clustalw/).

I have compared the performance of ULPB against both UJP and PSWM for all the binding sites in the *Escherichia coli K12 MG1655* for which I can obtain a training set of at least 20 sequences. I have also compared three binding sites in detail, including a comparison of the first order HMMs, and in two cases, ChIP-on-chip data. These are the cAMP-receptor protein (CRP), LexA and ArcA. *E. coli* has been chosen because of the large number of experimentally verified binding sites sequences available and so provides the best data to test these ideas. The known binding site training sequences in this study have been obtained from RegulonDB.
**CRP**: cAMP-receptor protein (CRP) is one of the seven ‘‘global’’ transcription factors in *E. coli* (Martinez-Antonio and Collado-Vides, 2003). It is known to regulate more than one hundred transcription units (Jacob and Monod, 1961). CRP’s activity is triggered by binding of the second messenger cAMP in response to glucose starvation and other stresses (Jacob and Monod, 1961). CRP binding sites have proved to be particularly noisy as the computational searching for the consensus-binding site can easily miss many known binding sites. CRP was chosen for its high promiscuity to the transcription factors.

**LexA**: LexA directly regulates ~30 *E. coli* transcription units involved in the "SOS" response (Walker, 2000). Such transcription is induced in response to DNA damage. Under normal growth conditions, LexA binds to a specific 20-base-pair (bp) sequence within the promoter regions of these genes, repressing transcription by sterically occluding RNA polymerase (RNAP). LexA was chosen for its lower promiscuity to the transcription factors, which should exhibit better behaviour than the CRP binding site.

**ArcA**: ArcA is a global regulator that changes in relation to the expression of fermentation genes and represses the aerobic pathways when *Escherichia coli* enter low oxygen growth conditions (Nikel et al., 2008). ArcA was chosen for its different protein domain (CheY like) and a very low consensus of the binding site.

## 2.4. Results

### 2.4.1 Neighbouring positions in binding sites show high levels of mutual information

I have identified base dependences in TFBSs using mutual information (Chouinard et al., 1996) between each base pair of the TFBS sequences. In all three binding sites analyzed (CRP, LexA and ArcA), there are high dependencies among the neighbouring positions (**Figure 2-2**) The CRP binding sites show high mutual information particularly between positions 5, 6, 7, 8 and between positions 15, 16, 17, 18 and19. These sites also show longer-range correlations between 6, 15, 17 and 19 and strong correlation between 7 and 16 and finally a correlation between position 8, 19 and 21. The LexA binding sites show higher correlations than the CRP binding site in most of the neighbouring positions. There are also many distant correlations e.g. in positions 5, 6, 7, 8 and 9 with the bases before 5 and position 5 with the bases 15 to 20. ArcA binding sites show multiple correlations in the distal and proximal five positions as well as some distant correlations between positions 3, 4 and 12, 13, 14 and 15. In all three cases, the central portion of the TFBS showed little mutual information between neighbouring bases. There are also frequent occurrences of distant correlated mutations in palindromic positions. Many transcription factors, including CRP and LexA, bind as dimmers (http://ecocyc.org/). Therefore, the associated transcription factor binding sites frequently consist of two similar anti-parallel sequences forming a separated, usually imperfect, palindrome.

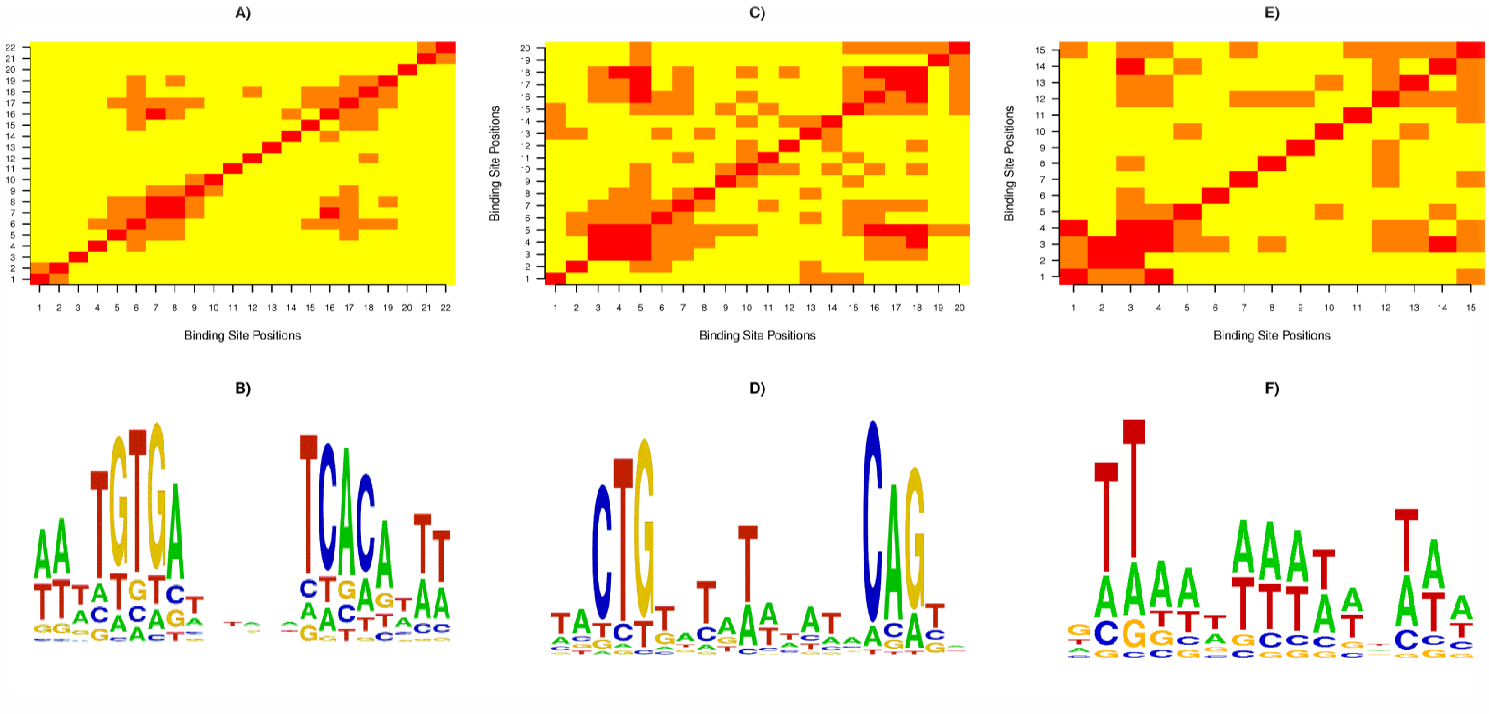


**Figure 2 - 2**: Heat maps and sequence logos (Crooks et al., 2004) of the three binding sites under study showing mutual information between bases. Darker squares indicate higher mutual information. (A,B) CRP, (C,D) LexA and (E,F) ArcA. For all three genes, there are high levels of mutual information between many neighbouring bases, as well as longer range interactions. Mutual information on the minor diagonal represents palindromic correlations.

Thus, correlations between the upstream and downstream portions of the TFBS are to be anticipated.

### 2.4.2 Predictions based on base inter-dependence outperform methods based only on position-specific information

I have assessed both the ungapped models and the HMM models for the three transcription factors mentioned using training set of their known binding sites .The Receiver Operating Characteristic (ROC) Curve (Zweig and Campbell, 1993) is shown for each binding site (**Figure 2-3**). For all three binding sites, the ULPB model shows a distinct advantage over other methodologies in predicting the binding sites.

The True Discovery Rate has been recorded for the binding sites tested against each method. The ULPB method shows a consistent improvement over position specific methods and other neighbouring based methods, with area under curve 0.97 for CRP, 0.88 for ArcA and 0.98 for LexA. This is compared with the PSWM giving 0.92 for CRP, 0.82 for ArcA and 0.82 for LexA. Without positional background, the ungapped likelihood performs marginally worse, with 0.95 for CRP, 0.87 for ArcA and 0.98 for LexA. The first order HMM shows a good performance for CRP with 0.96 and LexA with 0.98 On the other hand first order gapped HMM shows worst prediction for ArcA with 0.77 vs. 0.82 for normal ungapped position specific method (see discussion). Table (2-1) demonstrates the area under ROC curve calculations for all of the methods. The ULPB model performs at least as well or better than all other methods for all three binding sites analyzed in detail.

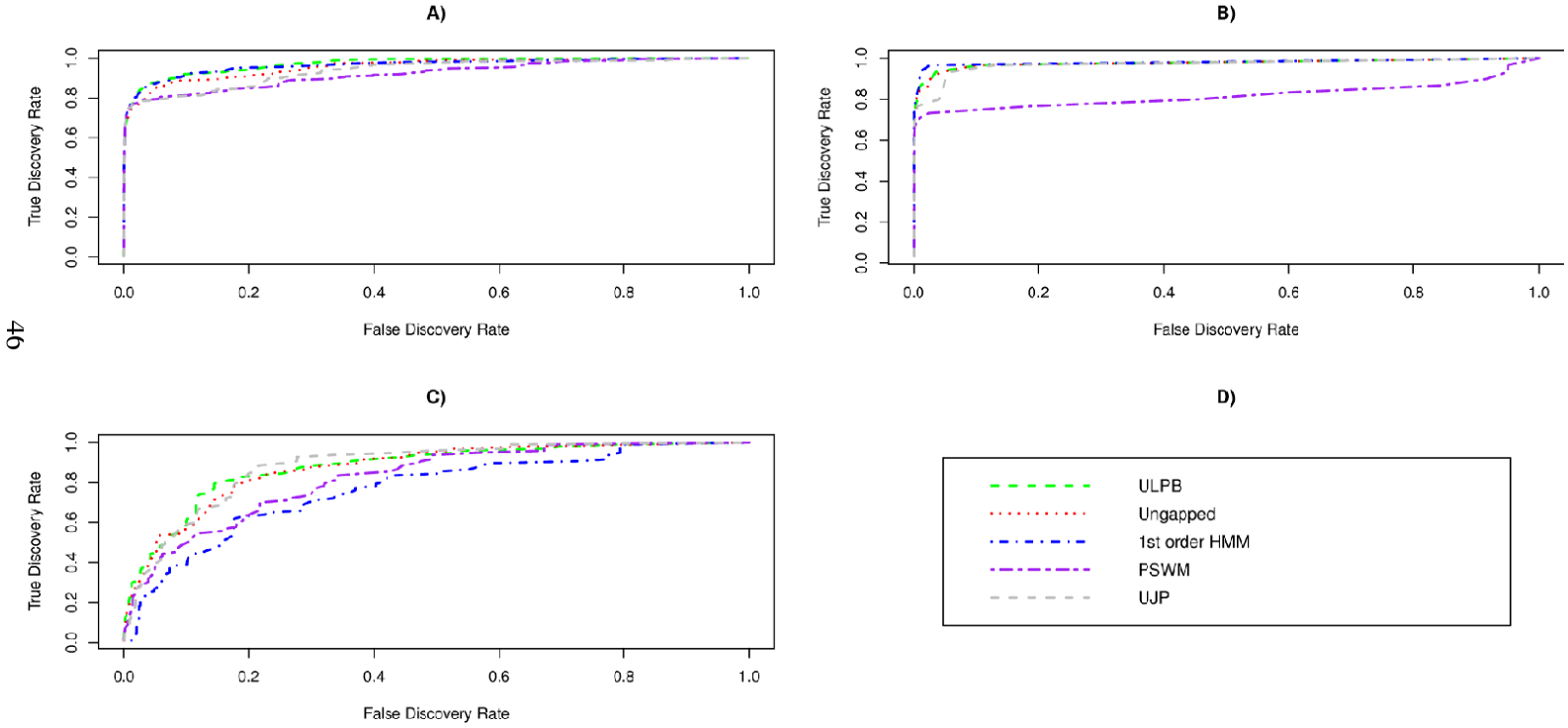


**Figure 2 - 3**: ROC curves for the binding sites being studied: A) CRP, B) LexA and C) ArcA. Each plot shows a comparison between Green: the Ungapped Likelihood under Positional Background, Blue: the gapped alignment scoring using Viterbi algorithm, Red: un-gapped alignment using the conditional probability, Purple: normal PSWM scoring and Grey: un-gapped joint probability. Observe that in all cases our novel ungapped method either outperforms or in same level of all other methods.

A thorough analysis of all 22 binding sites demonstrates that the ULPB model performs better than PSWM in every case see (Figure 2-4:6 and **Table 2-2**), with very substantial improvements in some cases (e.g. FlhDC). Generally, ULPB substantially outperforms the UJP method of Tomovic and Oakeley in at least 8 binding sites and is marginally better for further 8 binding sites with the same performance for the other binding sites (Figure 2-7). To assess the significance of the AUC enhancements, I have conducted two statistical tests using Wilcoxon Paired test between the AUC of ULPB vs. UJP and between the AUC ULPB vs. PSWM and corrected the p-values using Bonferroni correction. The statistical significance of the AUC between ULPB and UJP showed a significant p-value of 0.0009, while ULPB has shown a higher significance over PSWM (p-value < 1e-06).

Analyzing the performance of the method given the length of the binding site, we have found insignificant correlation between the AUC and the length of the binding site (p-value=0.054). A significant correlation is however found for UJP (AUC vs. length of the binding site) with p-value equal to 0.03, signifying the importance of using the conditional probability model in ULPB over the joint probability model used in UJP.

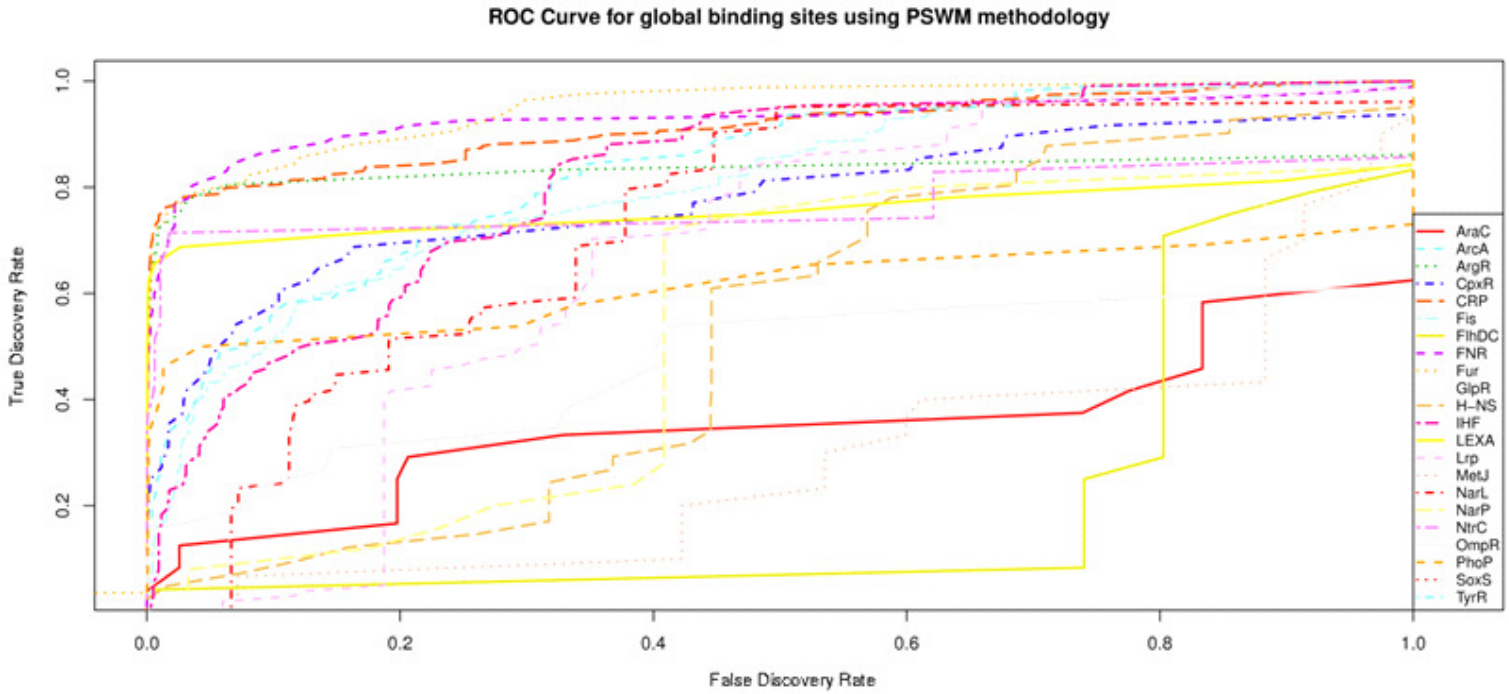


**Figure 2 - 4**: ROC curves for the 22 global binding sites being studied using Position Specific Weight Matrix.

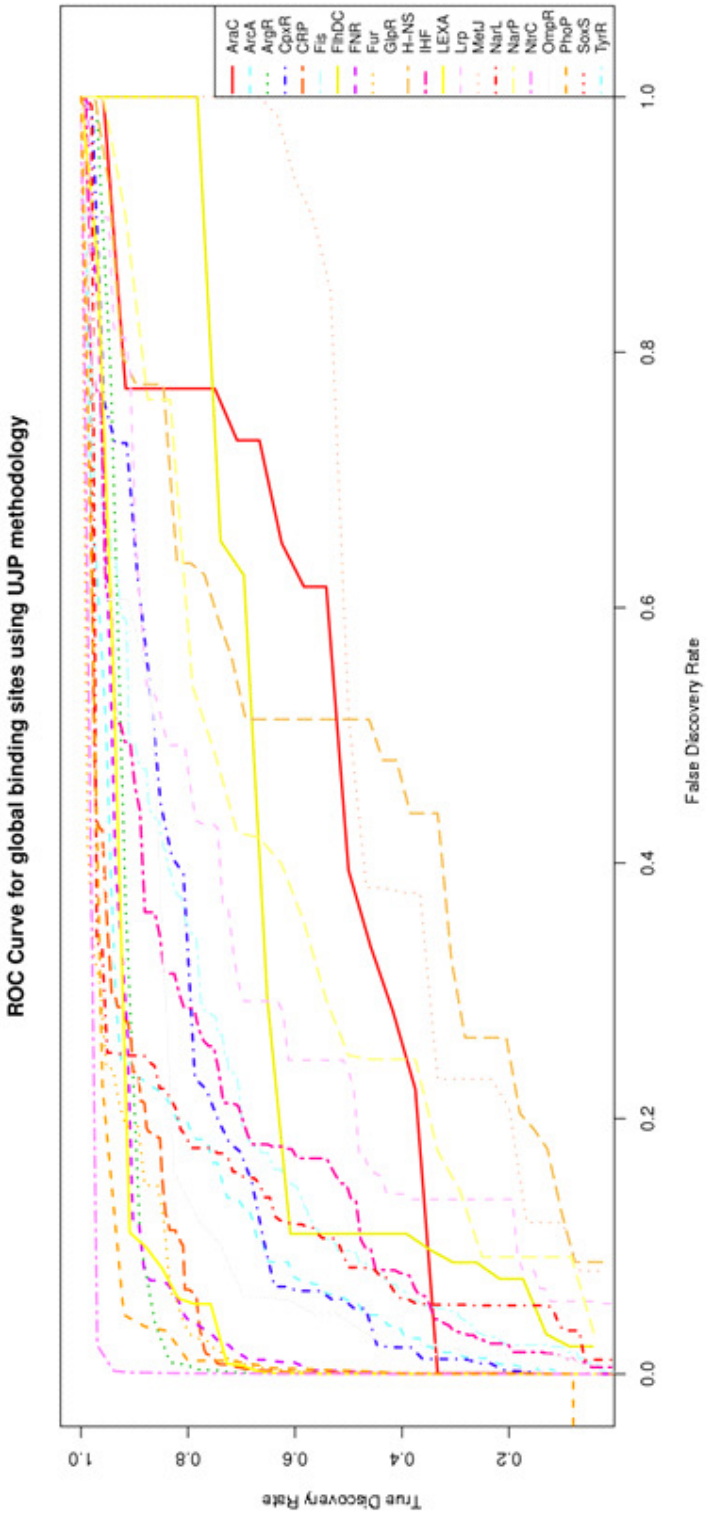


**Figure 2 - 5**: ROC curves for the 22 global binding sites being studied using Ungapped Joint Probability.

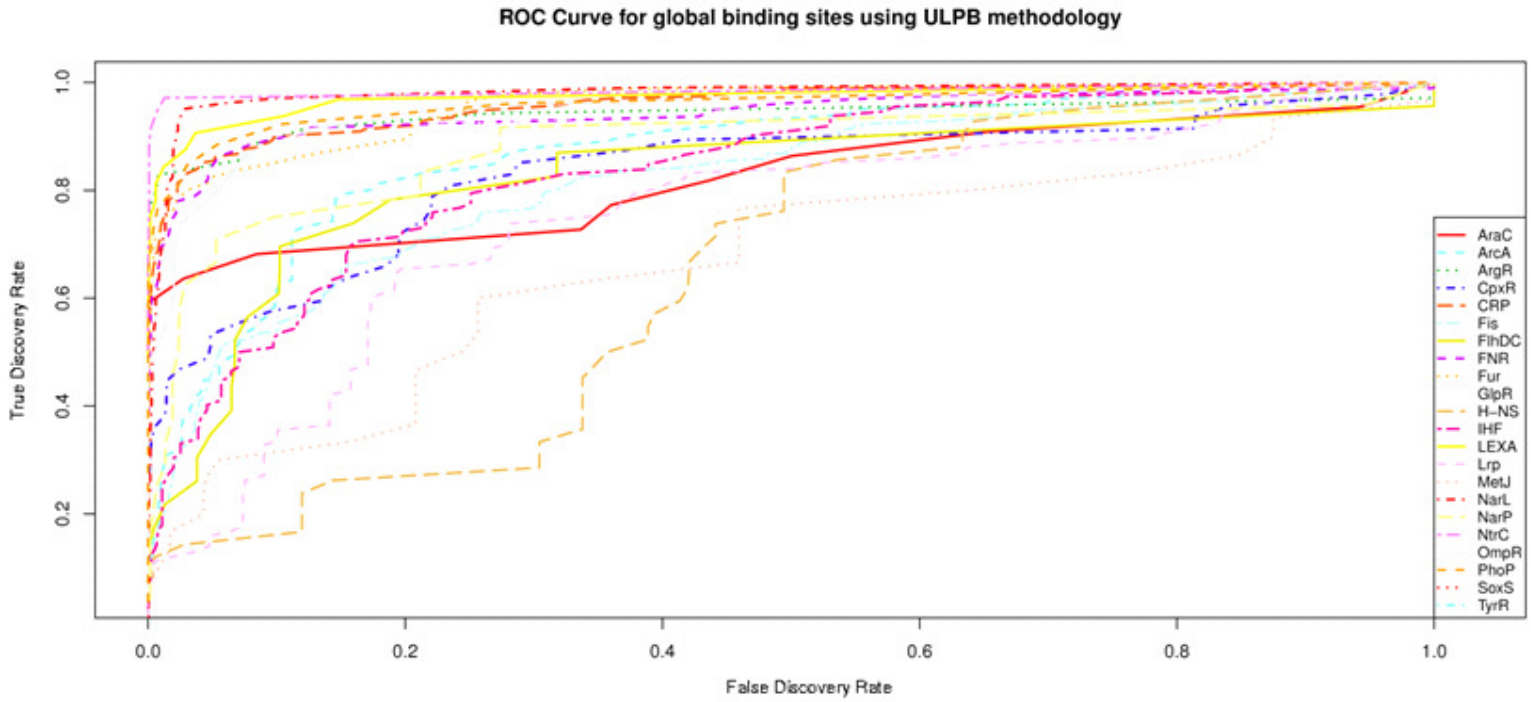


**Figure 2 - 6:** ROC curves for the 22 global binding sites being studied using Ungapped Likelihood under Positional Background**.**

| Binding Site | PSWM | First order HMM | UJP | Ungapped model | ULPB |
|---|---|---|---|---|---|
| CRP | 0.92 | 0.96 | 0.93 | 0.95 | 0.97 |
| LexA | 0.82 | 0.98 | 0.97 | 0.98 | 0.98 |
| ArcA | 0.82 | 0.77 | 0.88 | 0.87 | 0.88 |

**Table 2 - 1**: Area under ROC curves for all five methods applied to all three binding sites.

| Binding Site | Length | No of Sites | PSWM | UJP | ULPB |
|---|---|---|---|---|---|
| FlhDC | 16 | 20 | 0.22 | 0.67 | 0.85 |
| MetJ | 8 | 27 | 0.28 | 0.4 | 0.67 |
| AraC | 18 | 20 | 0.35 | 0.58 | 0.83 |
| OmpR | 20 | 22 | 0.47 | 0.87 | 0.96 |
| NarP | 7 | 22 | 0.55 | 0.64 | 0.88 |
| PhoP | 17 | 23 | 0.62 | 0.97 | 0.97 |
| GlpR | 20 | 23 | 0.61 | 0.91 | 0.93 |
| TyrR | 18 | 22 | 0.58 | 0.82 | 0.84 |
| LexA | 20 | 30 | 0.76 | 0.93 | 0.98 |
| NtrC | 17 | 32 | 0.78 | 0.99 | 0.99 |
| H-NS | 15 | 37 | 0.54 | 0.54 | 0.65 |
| ArgR | 18 | 33 | 0.83 | 0.92 | 0.95 |
| Lrp | 12 | 97 | 0.67 | 0.72 | 0.75 |
| SoxS | 18 | 20 | 0.86 | 0.58 | 0.95 |
| CpxR | 15 | 44 | 0.78 | 0.84 | 0.84 |
| ArcA | 15 | 97 | 0.82 | 0.88 | 0.88 |
| CRP | 22 | 275 | 0.92 | 0.93 | 0.96 |
| IHF | 13 | 106 | 0.82 | 0.83 | 0.84 |
| FNR | 14 | 91 | 0.93 | 0.93 | 0.95 |
| Fis | 15 | 245 | 0.81 | 0.82 | 0.82 |
| Fur | 19 | 81 | 0.95 | 0.96 | 0.96 |

**Table 2 - 2**: Area under ROC curves for two methods applied to all 22 regulators with at least 20 known binding sites in RegulonDB.

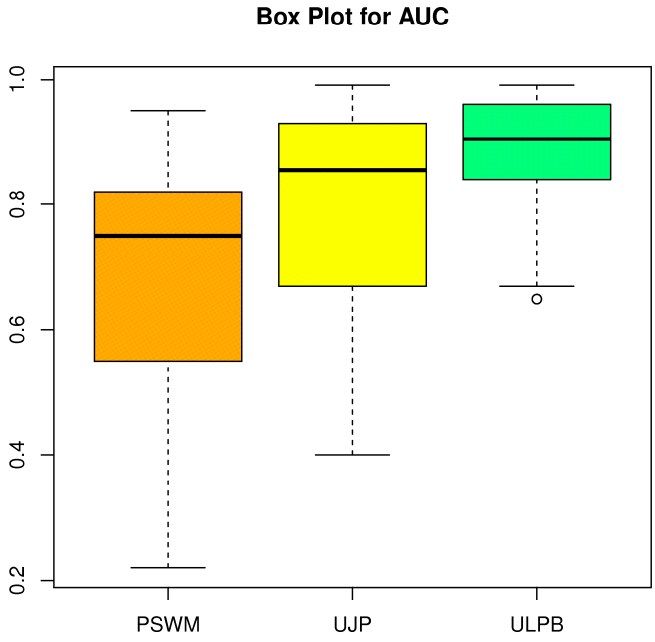


**Figure 2 - 7**: Box plot for the area under curves for all three methods compared indicating the enhancement of prediction given the algorithm.

### 2.4.3 ULPB Platform: A web interface to the ULPB methodology

ULPB is a website giving public access to the algorithm described in this chapter. It predicts binding sites from a set of search sequences based on a set of known binding sites sequences and using the ULPB method explained before. The website is integrated with xbase2 system (Choudhuri, 2004) giving user access to searching more than 600 bacterial genomes (as of august 2009).

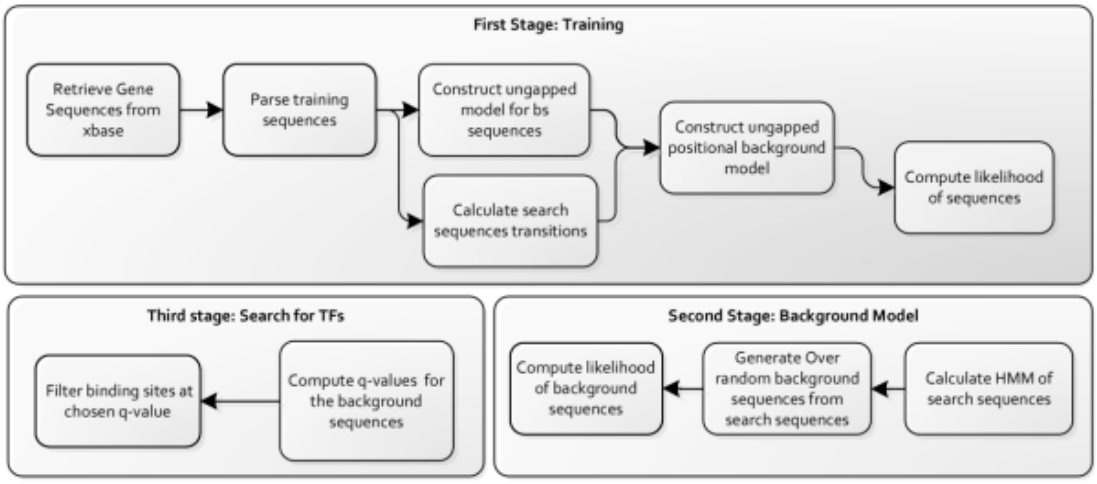


**Figure 2 - 8**: ULPB website passes through three stages in its process of the TFBS search. The First stage starts by computing the likelihood for the training sequences using ULPB. The Second stage, a background model is generated from the search sequences and is used as a null hypothesis. The Third stage determines the cut-off for the Transcription factor likelihood as 5% of the background sequences, and then it scores the given search sequences and outputs the binding sites over the 5 %.

The website searches in three stages (**Figure 2-8**): The first stage is training, the second stage is the background scoring, and third stage is for choosing the best cut-off for the binding site (default is 0.05 Q-value). A final option is motif filtering in which the returned predicted binding sites can be filtered by a user-supplied regular expression motif.

The binding site cut-off is selected with given q-value FDR cut-off under a set of background sequences generated from the search sequences given. The random set of sequences is generated after training a Markov chain on the transitions between the nucleotide types; in essence, a Markov chain is constructed as in (**Figure 2-1**).

The transition probabilities are captured from the search sequences. Starting with a random base [A, C, G, T], these probabilities are then used to generate a sequence with the same length as the binding site. The website is currently available on http://www.ulpb.bham.ac.uk

## 2.5. DISCUSSION

I have described a new methodology to score binding site likelihoods that uses interdependence between nucleotide bases and the positional background weights. I have shown that this provides a better scoring function compared to current position specific methodologies and existing base interdependent methods.

This method was tested in detail for three different *E. coli* transcription factors: CRP, LexA and ArcA, and against PSWM and UJP for a further 19 binding sites. *E. coli* was chosen, as it is the prokaryote for which the most number of representatives experimentally determined training sequences are available. CRP was chosen for its high promiscuity, ArcA was chosen for its low conservation and LexA was chosen for its high conservation. These chosen binding sites were chosen to represent most of the binding sites profiles. CRP and ArcA have shown a better performance on ULPB than the current position specific, hidden Markov model and UJP methods. LexA on the other hand has shown a close performance between the methods.

The binding sites studied are all global regulators. It is difficult to apply training-based methodologies for TFBS prediction for transcription factors that regulate only a small number of genes because these methods need an appropriately sized training set to generate a reliable model. One approach to get round this could be to build a training set using known transcription factor binding sites for homologous transcription factors in closely related organisms.

The ungapped likelihood under positional background method was better than the first ungapped model since it gives higher weight for the binding site-specific inter-dependencies

versus the background inter-dependencies, which increases the specificity of the method for the binding site against a certain set of search sequences. It has also shown improvement on binding sites prediction over the UJP method of Tomovic and Oakeley, which uses joint probabilities.

The ungapped methods presented here generally proved to be a better scorer than the Hidden Markov models that include gaps that are representative of insertion and deletion events. Thus, the interdependent effects are not as well captured by the evolutionary mutations included in a gapped MSA. In other words, the gapped alignment process actually disrupts the correlations between the bases forcing the HMM to select the best deletion or insertion or a nucleotide for the correlation, which introduces noise in the correlation profile. This effect is particularly apparent when comparing ArcA with LexA. The alignment introduces many more gaps in case of ArcA and almost no gaps with LexA (**Figure 2-9**), which could explain the difference between the blue curve (first order HMM) and the green curve (ungapped) in (**Figure 2-3**). Perhaps an alignment methodology that only allows internal gaps would perform better.

The mutual information analysis also revealed that binding sites sometimes exhibit palindromic correlations. However, a model that included correlations with palindromic positions was less successful than the ungapped model presented (data not shown). It is possible that a model that uses a graph-theoretic tour of the mutual information matrix to capture long range and palindromic correlations could be more successful (Barash Y, 2003, King and Roth, 2003).

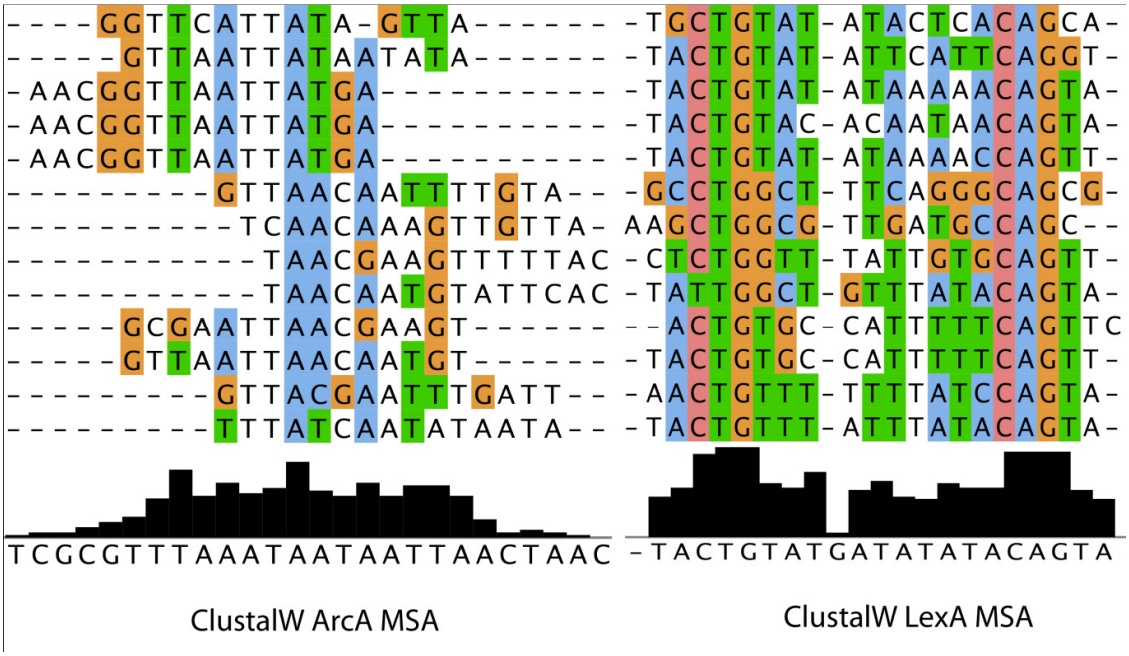


**Figure 2 - 9**: Jalview export of Clustalw MSAs for ArcA and LexA. The ArcA alignment has many gaps, especially at the start of the sequences. The LexA alignment has fewer gaps.

The combination of likelihood scoring and ChIP-on-chip or ChIP-seq analysis can be a powerful method for prediction of transcription factor binding sites. I have shown that for LexA, the ULPB method can help increase sensitivity of predictions without loss of specificity. The data for CRP were less conclusive; this is likely to be due to the high promiscuity of CRP for DNA suggesting that complex chemical interactions contribute to the binding, beyond the consensus of the binding site alone. Thus in principle the combination of computational and ChIP techniques is potentially effective, but care needs to be taken over choice of transcription factor for analysis.

The ULPB method although proving robust, have failed to predict some binding sites (MetJ, Lrp, HNS) to an acceptable accuracy. Analyzing those binding sites, we have found that the binding sites behaving worse are the ones characterized by at least two or more stable

consensus binding sequences rather allelic version of the same consensus sequence, where an approximation of both of them won't yield an optimal result.

A final possible extension of our work would be to use these methods to relate the sequence of the TFBS with its affinity. Such work would require a sizeable training set of measured affinities and could be useful in predicting the affinities of TFBSs for which no measurements are available.

# Chapter 3

# A THERMODYNAMIC APPROACH FOR ALIGNMENT OF NON-CODING AND NON-EVOLUTIONARILY-RELATED DNA SEQUENCES.

## 3.1. Abstract

Multiple Sequence Alignment (MSA) in general represents an important step in the evolutionary study of either protein or DNA sequences. The purpose of such a study is usually elucidating evolutionary clusters by assuming evolutionarily diverging insertion and deletion operators. Non-evolutionarily related sequences such as DNA non-coding binding sites for the same species cannot be aligned under the same hypothesis, hence another hypothesis must be drawn. In this work, I present a thermodynamic hypothesis to explain the variability among transcription factor binding sites (TFBS) in the same species. To verify such hypothesis I have tested two novel methodologies in aligning TFBS, first considering only base interdependence

hypothesis in the alignment and second by considering stacking free energies in the substitution matrix. I have assessed the prediction power of these two alignments for 18 of the global binding sites in *E. coli.* against each other and against other commonly used DNA alignment tools (ClustalW and Dialign). I have also devised a new alignment colouring scheme according to the base stacking free energy, which shows the possible thermodynamically alternative nucleotides with the same colour. I have demonstrated that considering free energies as an alternative hypothesis for multiple sequences alignment is superior to evolutionary hypothesis as most of the variability among binding site alleles can be explained.

## 3.2. Introduction

Multiple Sequence Alignment programs have been widely used to align protein or coding DNA sequences. These alignment programs rely on the evolutionary relatedness between sequences using substitutions, insertions and deletions as evolutionary operators to align them. These tools have proved extremely successful in their aim, structuring the relatedness between sequences and showing their high similarity regions.

Current alignment programs are mostly capable of aligning both protein and DNA coding sequences since the same evolutionary hypothesis holds for both of them. However, DNA non-coding regions for the same species, such as transcription factor binding sites (TFBSs), are not necessarily evolutionarily related, and frequently converge from non-common ancestors. Thus, the assumptions used to align protein sequences and DNA coding regions are inherently different from those that hold for non-coding sequences. While it is meaningful to align DNA coding regions for homologous sequences using mutation operators, alignment of binding site sequences for the same transcription factor cannot rely on mutation operators for an alignment. Similarly, the evolutionary operator of point mutations can be used to define an edit distance for coding sequences, but this has little meaning for non-coding sequences, since any sequence variations have to maintain a certain level of specificity for the binding site to function.

In this chapter, I aim to resolve these issues by designing multiple sequence alignment algorithms specifically optimized for non-coding DNA sequences, such as TFBSs. I will describe two methods, both of which align the binding sites using a thermodynamically inspired model for point/compensating mutations based on dinucleotide substitutions.

### 3.2.1 Dinucleotide Substitution Matrix

It has long been known that interactions between neighbouring DNA bases have a significant impact on DNA topology. For example, the thermodynamic properties of base stacking interactions have been extensively measured, and are commonly used in computational methods for DNA secondary structure prediction (Mathews et al., 1999). This was illustrated in work discussing the effect of DNA flexure on the binding site affinity (Calladine and Drew, 1986). Compensating mutations between neighbouring DNA bases have been long known (Stormo et al., 1986). I have also shown in Chapter 2 using mutual information analysis that there are dependencies between neighbouring and distant positions of the TFBSs that I studied, although the distant positions have been postulated to be a palindromic phenomenon.

Binding sites prediction has been shown to respond well to considering a dinucleotide approach, which could be due to stacking interactions. In Chapter 2, I have demonstrated that a model considering a simple block alignment of the binding site for the same transcription factor while representing them statistically as a dinucleotide position weight matrix enhanced the prediction power for most of the global binding sites in *E. coli.*

However, these approaches rely on a suitably large set of known binding sites and so predictions can only be made for "global" regulators. Most transcription factors only regulate a small number of genes and there are currently no methods that can predict binding sites for "non-global" regulators. One solution would be to use binding sites from paralogous or orthologous transcription factors, but in order to do so it becomes necessary to produce multiple sequence alignments of the binding sites. Interestingly, however, the predictions were

worsened when using sequence alignments based on existing MSA algorithms, as will be shown in chapter 4. Therefore, there is a clear need to develop appropriate MSA algorithms for non-coding DNA sequences.

In this chapter, I have used dinucleotides as the basis for multiple sequence alignments. I present two models for the calculation of the substitution matrices: the first is a statistical approach similar to the one used in BLOSUM (Henikoff and Henikoff, 1992); the second uses a Boltzmann distribution centred on the change in base stacking free energy as the null hypothesis for dinucleotide substitutions.

Both alignment methods are compared against each other and against other commonly used methods (Clustalw, Dialign). When testing protein MSAs, there are a number of benchmark alignments that are commonly used in protein alignment, including BAliBASE (Thompson et al., 2005) , OXBench (Raghava et al., 2003), SABmark (Walle et al., 2004) and SMART (Ponting et al., 1999). For DNA coding regions, Carroll *et al.* (Carroll et al., 2007) have developed a DNA reference benchmark based on the tertiary structure of encoded protein. However, no such benchmark alignments are available for non-coding DNA sequences. A score is however often used to detect the accuracy of the MSA using the homology in the resulting alignment as can be found in the work done by Thompson (Thompson et al., 1999). A better approach however to test the efficacy of the MSA algorithms is by testing their ability to predict TFBSs for global regulators in *E. coli* using a 1st order HMM and leave-one-out cross-validation, as described previously in Chapter 2 (Salama and Stekel, 2010). The predictive power of each alignment can then be assessed using area under Receiver Operating

Characteristic (ROC) curve (Zweig and Campbell, 1993); where this is a good measure of its efficacy.

## 3.3. Methods

### 3.3.1 Multiple Sequence alignment

The core of this work is not in the development of a new alignment process, but rather in the details of what is aligned and how the alignment is scored. Therefore I have adapted an existing alignment program, Opal (Wheeler and Kececioglu, 2007), which is highly configurable with accessible source code. Our alignment methodology makes profound changes in the alphabet, substitution matrix and gap penalties as explained below. I have optimized the specific alignment parameters in Opal to 1024 polish iterations, 3 trees and random three-two cuts.

### 3.3.2 Di Nucleotide Substitution Matrix

The Substitution Matrix considered in this case is *16×16* matrix which represents the dinucleotide substitution rate for the binding site. The dinucleotides are represented by assigning a new character for each pair of characters, using the alphabet A to P to represent alphabetically the dinucleotides AA through to TT. A DNA sequence is translated into our new alphabet in an overlapping manner, so that, for example, the sequence ACA would be represented in our new alphabet by the sequence BE: B for AC and E for CA. Every sequence in the training set is converted into this new alphabet for alignment with the other sequences.

The hypothesis behind this conversion is that the sequences are now forced into an inter-dependent representation of the binding site rather than an independent one. This representation

captures the heart of the single nucleotide mutation effect on neighbouring base interactions. For instance a single point mutation (transition) of the Cytosine to Thymine in this sequence ACA => ATA would result in a change of two neighbouring base interactions AC => AT and CA=>TA, and so would be represented as two position mutations in dinucleotide representation BE => DM. When considering stacking interactions, this can be used to represent the change in free energy of both interactions.

Since such a substitution matrix is highly specific for each binding site and they cannot generalize due to the specificity of each binding site, the substitution matrix is computed for each binding site individually. A problem is that to construct such a substitution matrix a valid alignment is required in the first place since you cannot rely on simple block alignment to represent the binding site consensus. This is inherently recursive, as obtaining a correct substitution matrix requires a valid multiple sequence alignment, and the multiple sequence alignment requires an optimized substitution matrix. This recursive nature of the problem have led us to devise a simple recursive Expectation Maximization like algorithm similar to the one devised by (Cao, 2009).

In the first step, the multiple sequence alignment is computed based on a predefined substitution matrix. In the maximization step, the alignment is pruned by removing the insertion columns (columns with more than 90% gaps), and the sequences with more than 30% Levenshtein distance (Levenshtein, 1966). The remaining alignment is then used to calculate the substitution matrix using equation (3-1). In the expectation step then, the resulting substitution matrix is used again to compute a new sequence alignment with optimum alignment cost. In all our implementations of this algorithm, it has converged; this is likely to

be because of the analogy with EM algorithms (Wu, 1983) but I have not attempted a formal proof of equivalence. The converged substitution matrix is then considered as a solution for this binding site, which is finally used to align the binding site using optimized gap penalties, as will be explained in alignment section.

The Cost matrix is constructed in general by using Log odds scoring techniques given by equation (3-1).

$$\mathrm{C}_{i,j} = K\left(1 - ln\frac{O_{i,j}}{E_{i,j}}\right) \quad \textbf{(3-1)}$$

Where $K$ is a scaling constant increasing the precision of the matrix, $O_{i,j}$ is the observed probability of aligning dinucleotide *i* with dinucleotide *j*, $E_{i,j}$ is the expected probability (Null Hypothesis) of aligning dinucleotide *i* with dinucleotide *j*.

Reversing the sign of the odd score function is pure technicality of the alignment program, which requires low cost values for expected substitutions and high cost for unexpected ones. Hence reversing the sign would penalize the negative odd score and reward the positive ones.

The observed probability $O_{i,j}$ is constructed using the following equation

$$O_{i,j} = \begin{cases} \sum_{x=1}^{n} N_i^x N_j^x, & if\ i \neq j \\ \sum_{y=N_{i-1}^x}^{1} y, & if\ i = j \end{cases} \quad \textbf{(3-2)}$$

Where $N_i^x$ is the number of dinucleotides *i* at position *x*

In this chapter, I have used two different Null hypothesis distributions, statistical and thermo dynamical, each of which result in two different cost matrices which in turn result in two

different alignment methodologies. These have been denoted as SDNMSA for the statistically generated null hypothesis and EDNA for the thermodynamically generated null hypothesis.

### 3.3.2.1. Statistically generated Null Hypothesis (SDNMSA)

The first null hypothesis given by equation (3-5) is a purely statistical hypothesis representing the expected independent joint distribution of the dinucleotides which is generated following the equations as given in (Henikoff and Henikoff, 1992).

$$Q_{i,j} = \frac{O_{i,j}}{\sum_{i=1}^{N}\sum_{j}^{i} S_{i,j}} \quad \textbf{(3-3)}$$

$$E_i = Q_{i,i} + \sum_{i \neq j}^{N} {}^{Q_{i,j}}/_{2} \quad \textbf{(3-4)}$$

Where $E_i$ is the expected probability dinucleotide *i*

$$E_{i,j} = \begin{cases} E_i E_i, & if\ i = j \\ 2E_i E_j\ , & if\ i \neq j \end{cases} \quad \textbf{(3-5)}$$

From equation (3-1), a statistical cost matrix is then generated using

$$S_{i,j} = K\left(1 - ln\frac{O_{i,j}}{E_{i,j}}\right) \quad \textbf{(3-6)}$$

Where $S_{i,j}$ is the statistical cost of substituting dinucleotide *i* with dinucleotide *j*, *K* is equal to 100.

### 3.3.2.2. Thermodynamically generated Null Hypothesis (EDNA):

The second null hypothesis is thermo dynamical and assumes an energetically independent alignment using the Boltzmann distribution derived from the dinucleotide stacking free energy (Allawi and SantaLucia, 1997, SantaLucia and Turner, 1997, Allawi and SantaLucia, 1998c, Allawi and SantaLucia, 1998a, Allawi and SantaLucia, 1998d, Allawi and SantaLucia, 1998b). The null hypothesis distribution is computed as follows; 1) Create a Boltzmann distribution of the existing dinucleotides in the training set using equation (3-7), 2) Create a joint distribution of the Boltzmann distribution assuming independence.

$$B_i = \frac{d_i \, e^{-\Delta G_i / K_B T}}{\sum_{i=1}^{16} d_i \, e^{-\Delta G_i / K_B T}} \qquad \textbf{(3-7)}$$

Where $B_i$ is the Boltzmann distribution of dinucleotide i, $\Delta G$ is the stacking free energy of the dinulceotide *i,* $T$ is the room temperature 295 Kelvin and $K_B$ is the Boltzmann Constant 0.00198721 kcal/mol/K

From equation (3-1), the thermo dynamical cost is generated as follows:

$$E_{i,j} = K(1 - ln\frac{S_{i,j}}{B_i B_j}) \qquad \textbf{(3-8)}$$

Where $E_{i,j}$ is the thermodynamic cost of substituting dinucleotide *i* with dinucleotide *j.*

The second null hypothesis represents the probability that two dinucleotides would align thermodynamically by chance.

The log odd scoring provides a score for the odds of the observed distribution assuming dependence versus either of the expected null hypothesis distribution. The two null hypotheses in comparison are profoundly different from each other, and one of them should provide a more approximation to the expected distribution.

### 3.3.3 Gap Penalties

The gap penalty function used is affine gap penalty function where I use two sets of gap penalties, one set for gaps within the sequence (internal) and another set for terminal or prefix and suffix gaps. Each set is two gaps, one for the initial gap and a another one for extending this gap, so I end up with four gap penalties ($\gamma, \beta, \gamma', \beta'$) as follows in equation (3-9)

$$G = \sum_{i=1}^{n}(\gamma + \beta D) + \sum_{j=1}^{2}(\gamma' + \beta' D') \quad \textbf{(3-9)}$$

Where $G$ is the total gap penalty for a certain sequence, $i$ is the internal gap regions, $j$ is the terminal gap regions (2 regions in this case, one at start and the other at the end), $\gamma$ is the internal gap open penalty, $\beta$ is the internal gap extension penalty, $D$ is the length of the internal gap extension including the opened one, and $\gamma'$ is the terminal (external) gap open penalty, $\beta'$ is the terminal gap extension penalty, $D'$ is the length of the terminal gap extension including the opened one.

The gap penalty has to be closely related to the cost matrix, since an extremely higher gap than the cost would result in no gaps and an extremely lower gap penalty would result in long gaps in the alignment. Accordingly, the gap penalty values used in the alignments must be relative to

the cost matrix values. Therefore, the gap penalties used are taken as weighted averages of the cost matrix as follows in equation (3-10).

$$G = \sum_{i=1}^{n}(\gamma W + \beta WD) + \sum_{j=1}^{2}(\gamma' W + \beta' WD') \quad \textbf{(3-10)}$$

Where $W$ is the average of the cost matrix optimized previously and gap penalties used are factors of $W$.

Since our alignment mainly operates by shifting the binding sites versus each other, rather than using internal gaps, due to the steric constraints of protein-DNA interactions, so the internal gaps $(\gamma, \beta)$ open/extension are fixed at a high gap penalty factor of **5**. The terminal open/extension gaps are optimized where $\gamma'$ is fixed at 0.1 while the extension terminal gap penalty $(\beta')$ is optimized using an incremental search algorithm from 0.11 to a value where no gaps can be added to the sequence. In other words, the opening of a terminal gap is always allowed, while the extension is optimized.

The acceptance function for the penalty is optimizing two objectives; first) minimizing the length of the alignment, and second) maximizing the number of columns $\varphi$ with similarity percentage $(\propto)$ across the alignment columns as given by equation (3-11).

$$\varphi = count\left(\frac{\max(P_i)}{N} \geq \propto\right) \quad \textbf{(3-11)}$$

Where $\varphi$ is the number of columns in the alignment with similarity percentage greater than $\propto$ and $P_i$ is the number of similar bases at column i.

The homology percentage $\propto$ is chosen as the maximum percentage to obtain at least one column ($\varphi \geq 1$) in the block alignment ($P_B$) as shown in equation (3-12).

$$\propto = argmax_{\propto}\ \varphi(\propto, P_B) \qquad \text{, where } \varphi \geq 1 \tag{3-12}$$

Where $\propto$ is maximum possible $\propto$ for block alignment $P_B$ with $\emptyset \geq 1$

Accordingly; the optimum value chosen for $\beta'$ is the one corresponding to maximum $\epsilon$ as defined in equation (3-13) and described in (Figure 3-1) for AraC.

$$\varepsilon = \varphi/(L_A - L_B) \tag{3-13}$$

Where $L_A$ is the length of the alignment and $L_B$ is the initial length of the binding site block alignment (ungapped).

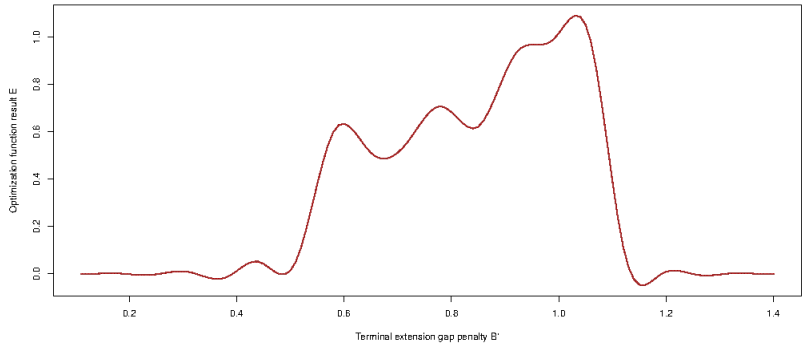


**Figure 3 - 1**: Terminal extension gap penalty $\beta'$ optimized against ε for AraC binding site, choosing the optimum penalty for the alignment that maximizes the optimization function; in this case the chosen terminal extension gap penalty is 1(i.e. average weight of the substitution matrix) which results in the maximum values for ε.

### 3.3.4 Multiple Sequence Alignment assessment through 1st order HMM TFBS Prediction

The assessment of the binding site alignment is done through the analysis of the transcription factor prediction sensitivity and specificity for 18 of the global regulators in *E. coli* K12. I have compared four alignment methods, ClustalW (Thompson et al., 2002), Dialign (Morgenstern, 2007), SDNMSA and EDNA. ClustalW was chosen as one of the most heavily used alignment tools for DNA alignment; Dialign has proved to be particularly successful in aligning DNA sequences (Morgenstern, 2007).

The assessment of the four alignment tools proceeded as follows:

1. The binding site is aligned using each of the alignment methodologies.

2. The training set likelihood is evaluated using a cross-out technique, by training the HMM model using N-1 binding sites and obtaining the likelihood for the remaining one.

3. A null hypothesis distribution is generated by training the first order HMM using the full training set of the binding site and scoring the likelihood over an overlapping window of binding sites for 200 base pair upstream of genes in *E. coli MG1655.*

4. A p-value is assigned for each binding site likelihood (obtained in step 2), versus the likelihood distribution of the null hypothesis.

5. p-values are then corrected for False Discovery Rate (Hochberg, 1995).

6. The p-values obtained are then assessed in a Receiver Operating Characteristic (ROC) Curve for every binding site and alignment methodology.

7. The area under each ROC curve is calculated and given in (Table 3-1).

### 3.3.5 Multiple Sequence alignment thermo dynamical colouring

The dinucleotide alignment was coloured using the following steps:

1. Convert the binding site alignment into the dinucleotide representation.

2. Cluster the dinucleotides based on free energies to group similar dinucleotides together.

3. Colour the clustered dinucleotides with the same colour if they belong to the same cluster.

4. Convert the dinucleotide to single nucleotide representation keeping the colour of the initial nucleotide of each dinucleotide.

## 3.4. Results

### 3.4.1 Dinucleotide alignment better conserves binding site sequences

Dinucleotide alignment colouring coded by thermodynamic clusters has shown that using such an alignment increases the number of positions that are energetically close to each other, and only shifts the sequences if needed. In addition, it has shown that the thermodynamic alignment provides a better performance in decreasing the positions with high free energy. On the other hand, single nucleotide alignment methods have shown such defects, as ClustalW for example showing a high free energy position in the middle and missing the conserved low free energy positions shown by proposed methods (**Figure 3-2**). Dialign provided a smaller conserved motif leaving out many conserved positions in the binding site as shown in (**Figure 3-2**).

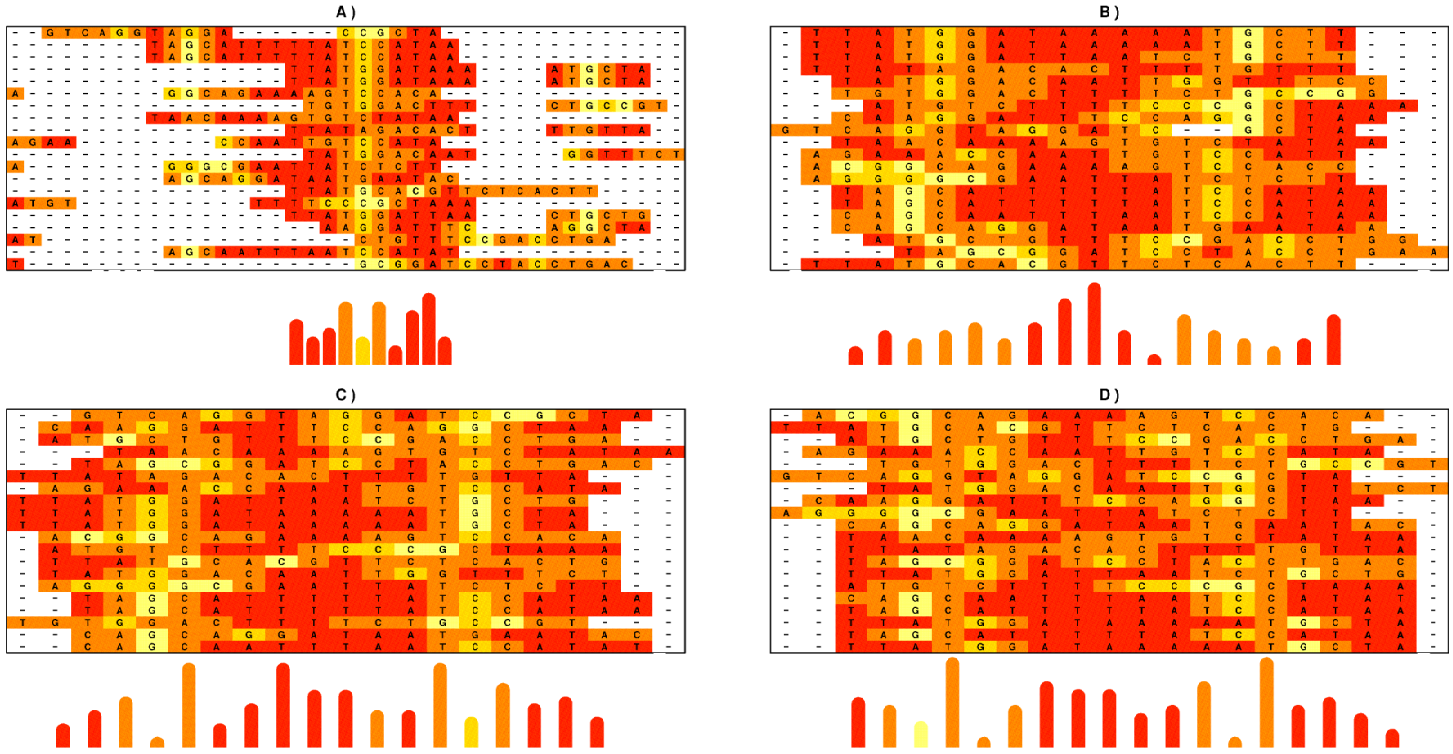


**Figure 3 - 2**: Alignment of AraC as an example using the four alignment methods introduced in the chapter. A) Dialign, B) ClustalW, C) SDNMSA, D) EDNA. The colours used follows a heat colour palette representing red as the highest free energy cluster and yellow as the lowest free energy cluster.

### 3.4.2 Dinucleotide representation of binding sites provides a better optimality of the alignment.

The dinucleotide representation of the binding site have proved to produce better alignments, providing better sensitivity and specificity of transcription factor prediction than either ClustalW or Dialign ( **Figure 3-3, Figure 3-4, Figure 3-5 and Figure 3-6**). In most cases, ClustalW have been found to behave better than Dialign (Table 3-1) but also behaved worse in other cases. On the other hand, the assessment shows consistent superiority of up to 70% enhancement using either of the new dinucleotide based alignments, particularly when the binding site sequences are less well conserved. The performances are similar for highly conserved binding sites as ArgR, LexA, CRP, GlpR and PhoP, which is expected since none of the alignments introduces much deviation from the block alignments.

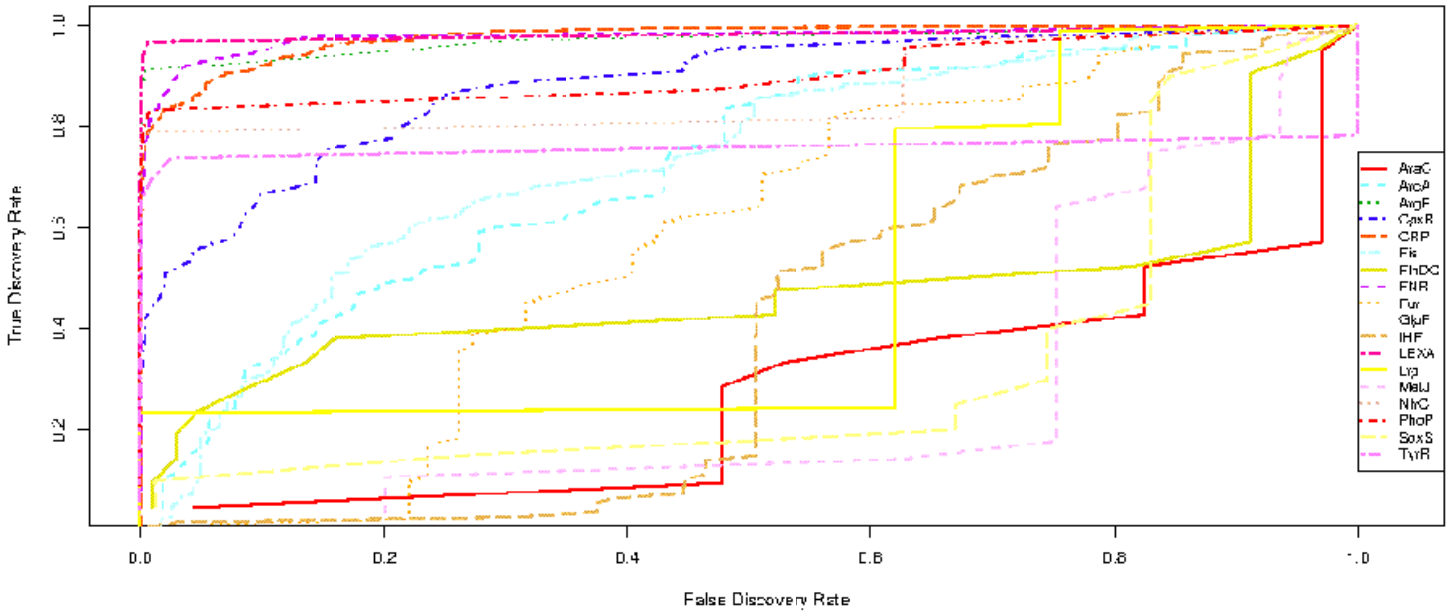


**Figure 3 - 3**: ROC Curve of 1st order HMM prediction of 18 *E.coli MG1655* binding sites using Dialign alignment

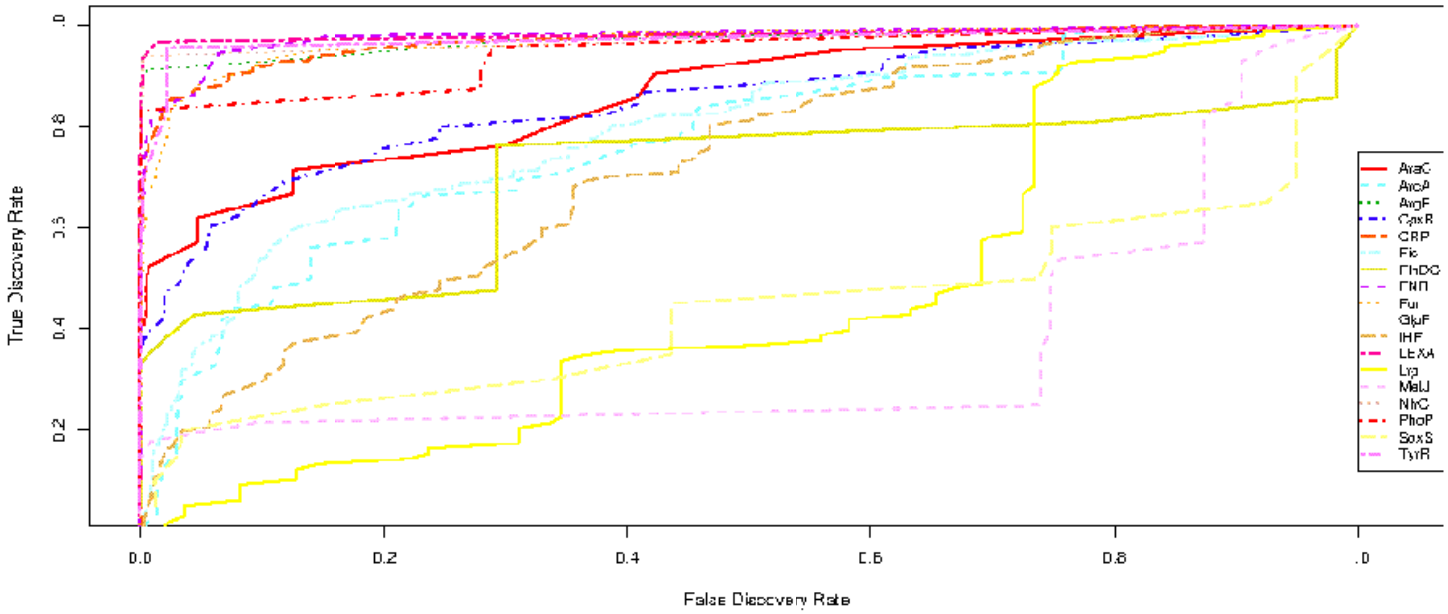


**Figure 3 - 4**: ROC Curve of 1st order HMM prediction of 18 *E.coli MG1655* binding sites using ClustalW alignment

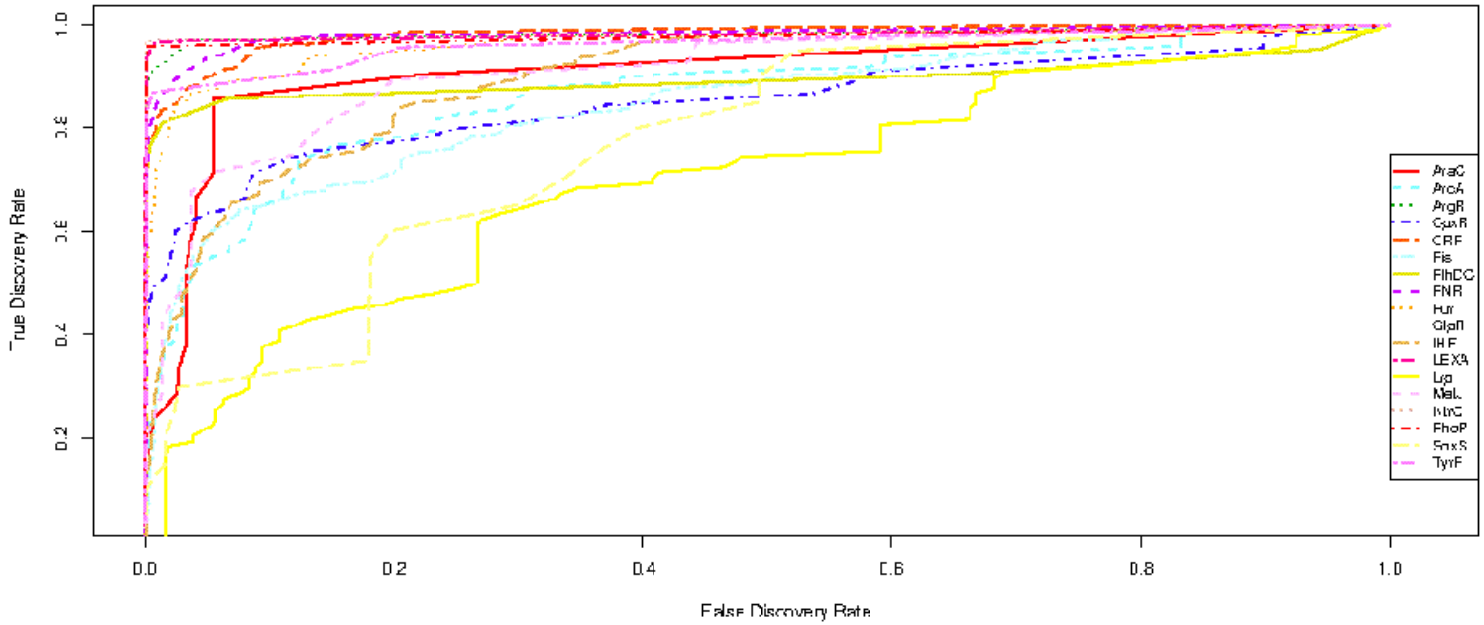


**Figure 3 - 5**: ROC Curve of 1st order HMM prediction of 18 *E.coli MG1655* binding sites using dinucleotide alignment based on statistically computed null hypothesis distribution

### 3.4.3 Thermodynamic based alignment proved better than simple statistical null hypothesis

The joint Boltzmann distribution null hypothesis has been found to enhance alignment optimality in most binding sites with up to 10% enhancement over the statistical null hypothesis. The thermodynamic alignment behaved better in 50% of the cases tested (AraC, ArcA, CpxR, Fis, FlhDC, Fur, IHF, SoxS, MetJ, NtrC and Lrp) and similar in 40% of the cases as shown in (**Figure 3-6**) and (**Table 3-1**).

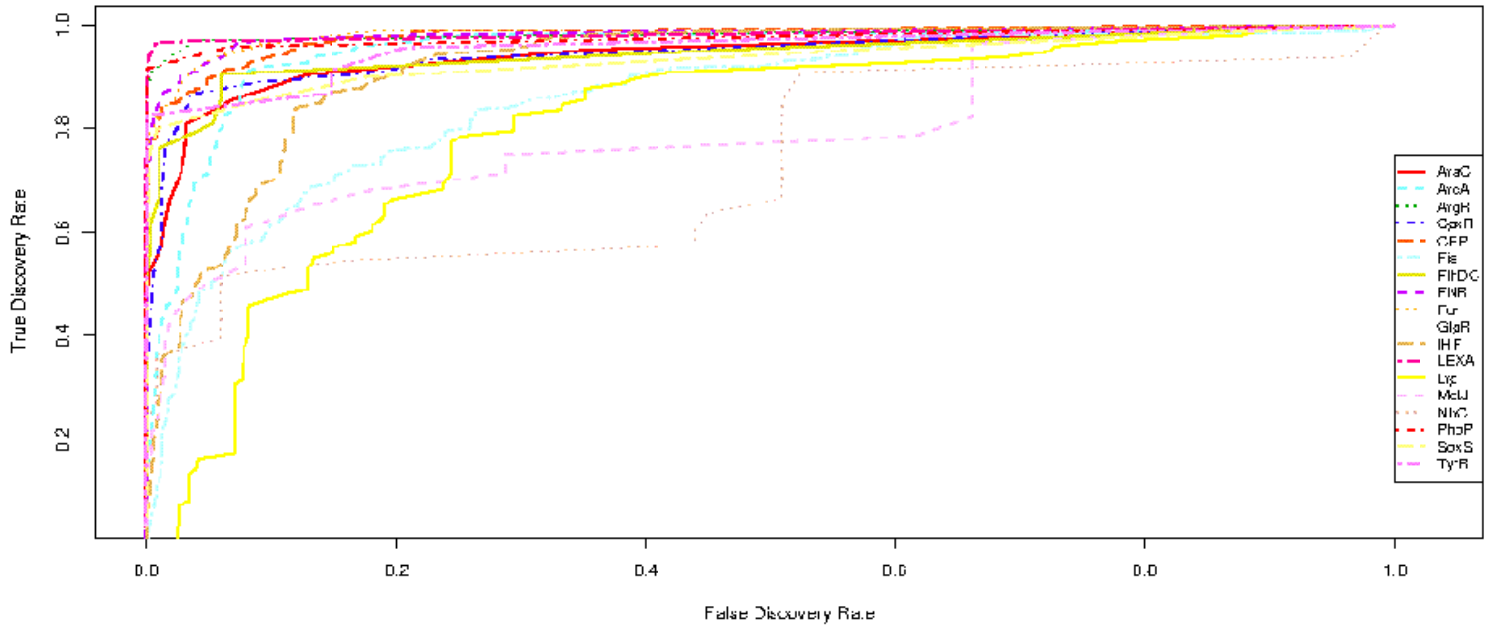


**Figure 3 - 6**: ROC Curve of 1st order HMM prediction of 18 *E.coli MG1655* binding sites using dinucleotide alignment based

### 3.4.4 1st order HMM based on thermodynamic dinucleotide based alignments outperforms block alignment based methods

The 1st order HMM described in our previous work behaved worse than block alignment prediction method, ULPB (Salama and Stekel, 2010). I have found that the first order HMM using either the statistical or the thermo dynamical alignment provides a consistently better behaviour than ULPB and provide similar prediction power for highly conserved binding sites (**Table 3-1**).

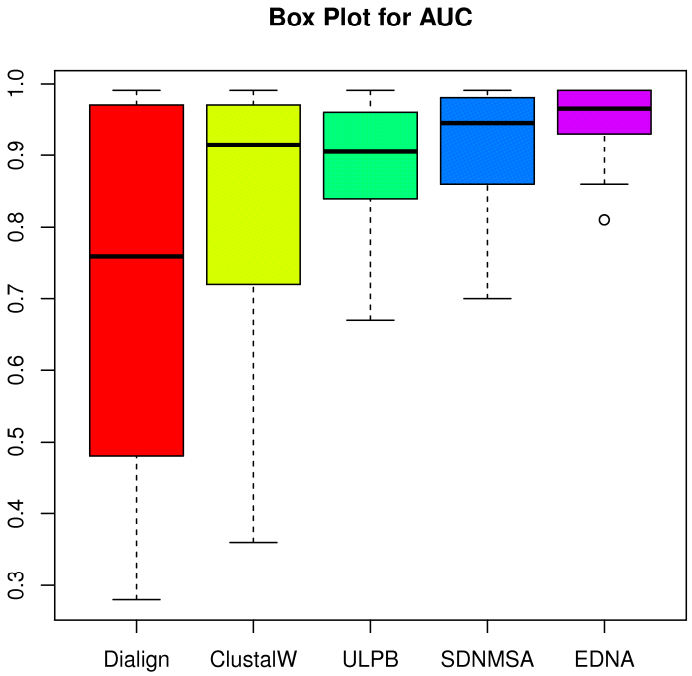


**Figure 3 - 7**: Box plot for the area under curves for all five methods compared indicating the significance of the prediction power for the corresponding method.

A statistical significance of the AUC was tested for all methods using a paired Wilcoxon test. It has shown the superiority of EDNA over other methods after Bonferroni correction (Dialign: 0.004388202, ClustalW: 0.003036702, ULPB: 0.002164233, SDNMSA: 0.022917506). The significance of EDNA over all other methods even SDNMSA shows the significance of using the thermodynamic driven expected distribution underlying the thermodynamic nature of the observed distribution.

| | Dialign | ClustalW | ULPB | SDNMSA | EDNA |
|---|---|---|---|---|---|
| AraC | 0.29 | 0.88 | 0.83 | 0.93 | 0.96 |
| ArcA | 0.72 | 0.77 | 0.88 | 0.86 | 0.95 |
| ArgR | 0.97 | 0.97 | 0.95 | 0.99 | 0.99 |
| CpxR | 0.89 | 0.86 | 0.84 | 0.86 | 0.95 |
| CRP | 0.98 | 0.97 | 0.96 | 0.98 | 0.98 |
| Fis | 0.74 | 0.79 | 0.82 | 0.85 | 0.86 |
| FlhDC | 0.48 | 0.71 | 0.85 | 0.91 | 0.96 |
| FNR | 0.98 | 0.98 | 0.95 | 0.98 | 0.98 |
| Fur | 0.58 | 0.97 | 0.96 | 0.96 | 0.99 |
| GlpR | 0.98 | 0.97 | 0.93 | 0.97 | 0.98 |
| IHF | 0.40 | 0.72 | 0.84 | 0.90 | 0.93 |
| LEXA | 0.99 | 0.99 | 0.98 | 0.99 | 0.99 |
| Lrp | 0.50 | 0.46 | 0.75 | 0.70 | 0.81 |
| MetJ | 0.28 | 0.36 | 0.67 | 0.91 | 0.91 |
| NtrC | 0.88 | 0.98 | 0.99 | 0.99 | 0.99 |

| | | | | | |
|---|---|---|---|---|---|
| PhoP | 0.91 | 0.95 | 0.97 | 0.99 | 0.99 |
| SoxS | 0.32 | 0.44 | 0.95 | 0.77 | 0.93 |
| TyrR | 0.78 | 0.98 | 0.84 | 0.97 | 0.97 |

**Table 3 - 1**: Table listing the Area under ROC curves for prediction sensitivity and specificity corresponding to 4 alignment methods (Dialign, ClustalW, SDNMSA, EDNA) applied to 18 of the global regulators with at least 20 known binding sites in RegulonDB *E. coli* MG1655, along with the simple block alignment based prediction ULPB (chapter 2).

## 3.5. Discussion

Consensus of the binding site can be defined as the conservation of the DNA binding site nucleotides resulting in binding motifs. The problem with such a definition is that for the same transcription factor, there can be a huge variability in the recognition of binding sites. This binding site consensus variability can explain the degrees of specificity for each binding site, although the puzzle gets more complicated as what are the limits to the variability of the binding site, and what are the governing rules to limit such variability. A simple null hypothesis would consider base interdependence to define the variability in consensus; in other words, "a nucleotide variation is allowed if it doesn't affect the interdependence relation between neighbouring nucleotides".

Accordingly, I have developed a new methodology for multiple sequence alignments of non-coding DNA sequences that uses an alignment of dinucleotides. To test the efficiency of the method, two variants have been described, using two different null hypotheses, one statistically driven and the other thermodynamically driven. These have been compared with other

alignment programs, exemplified by ClustalW and Dialign, which are designed with a null hypothesis derived from the evolution of peptide sequences. They have been compared using the transcription factor binding sites of 18 global regulators from *E. coli* K12. The ROC curves for the first order HMM prediction using the dinucleotide alignment demonstrated better alignment than the current alignment tools, irrespective of the null hypothesis used. Between the two variants, the use of thermodynamic driven null hypothesis proved to be statistically better.

In general, driving a Boltzmann distribution of the dinucleotides provides a direct distribution of the dinucleotides in a binding site based on the stacking free energy. An independent thermodynamic joint distribution would then represent the distribution of thermodynamically aligning two dinucleotides by chance. The log odd scoring system in this case provides a score of zero or negative if the two dinucleotides align thermodynamically by chance and are not expected to align by chance in case of a negative score. A positive score in this case would present that the null hypothesis provide a lower joint probability than the observed probability, which would be prove that such two dinucleotides are not aligning by thermodynamic chance. So scoring the observed joint distribution versus this null hypothesis distribution indicates how far dinucleotides pair is from independently aligning. In other words, the presence of dinucleotide $x$ in position $i$ of binding site $k$ can be substituted to a high probability by dinucleotide $y$ in position $i$ of binding site $k+1$ which can still provide stable energetic configuration for the same transcription factor to bind.

### 3.5.1 Thermodynamic null hypothesis is not governed by rarity of the dinucleotide

Another major advantage of using a thermodynamically driven null hypothesis is that it is not governed by the rarity of the substitutions in the binding site training set as pointed out in (Eddy, 2004), but rather provides a constant behaviour for all the binding sites, believing that the rarity of an alignment event is completely governed by stacking free energy. Thus, it avoids problems resulting from under-sampling in a set of transcription factor binding sites, which may be particularly valuable when building alignments for non-global regulators.

### 3.5.2 Base stacking interaction driven convergence

In general, the method provides a good performance for most binding sites relative to alternative methods. Nevertheless, false positives are still observed in some of binding sites, including Lrp, SoxS, IHF and MetJ. For very well conserved binding sites, such as LexA, ArgR, FNR, Fur, GlpR, PhoP, all methods show good performance; this can be attributed to the small solution space providing a set of constraint alignment solutions. On the other hand, the predictions for the less conserved binding sites, such as AraC, ArcA, CpxR, FlhDC, TyrR and MetJ, have been particularly enhanced by our new method. This may reflect the prevalent thermodynamic interactions, which have been captured well by this method.

The previous argument suggests that variability cannot be completely explained by base stacking interaction driven convergence but can be explained largely in some of the binding

sites where high stacking interactions are expected to prevail and overrule other converging factors.

### 3.5.3 Higher order matrices

The fact that I have used a first order substitution matrix to capture the binding site might be on its own providing an advantage for the alignment, since I am providing finer details for the substitutions and aligning blocks of two bases rather one. One may well argue, if the case would be better for N-order matrices. The hypothesis behind using a dinucleotide matrix is the stacking free energy, which is believed to cascade so the stacking free energy of N-nucleotides can be simply computed from the dinucleotides involved as no forces have been found to exist beyond the neighbouring base. Also the problem with such matrices is the complexity involved, for instance using a Tri-Nucleotide matrix would require a substitution matrix of $4^3 \times 4^3$, which would be immense to compute and optimize. Also the stacking free energies used have only been experimentally measured for $1^{st}$ order interactions and not for higher order interactions (Allawi and SantaLucia, 1997, SantaLucia and Turner, 1997, Allawi and SantaLucia, 1998c, Allawi and SantaLucia, 1998a, Allawi and SantaLucia, 1998d, Allawi and SantaLucia, 1998b). Also statistically considering a higher order interaction would suffer being hugely under sampled resulting in a lower statistical power. Finally the longer the combinations of nucleotides in alignment, the more restrictive would the search space be for an optimal alignment, since I am shifting larger blocks together.

### 3.5.4 Interspecies variability

Although this method hasn't been tested with binding sites collected from other species, the work done by (Moses et al., 2003, Moses et al., 2006) has explained that rate of variability among species for the binding site is position specific, identifying some positions as functionally important positions where the rate of variation is much slower than the other position, suggesting a heavier selection of such positions. The hypothesis presented in Chapter 2 suggests that slower variability positions are mainly involved in direct binding of transcription factor amino acids to nucleotides and hence not affected directly by stacking interactions. Accordingly; deviation in those direct interaction positions might result in loss of binding site specificity, while the higher variability positions would be ideally constrained by stacking interaction conservations while still showing variations both inter and intra species. Although such a hypothesis would need to be tested, the conclusion is still compatible with the work done.

# Chapter 4

# COMBINING LIKELIHOOD WITH CHIP-ON-CHIP SIGNAL CAN IMPROVE PREDICTION

## 4.1. Abstract

The challenging nature of binding site prediction has led to a considerable level of research in this area. This prediction is a non-trivial task because transcription factors are often promiscuous, binding to binding sites exhibiting varying patterns of nucleotide sequences. ChIP-on-chip provides a high throughput method of finding the sites where a transcription factor binds to chromosomal DNA. Although this method provides a list of prospective locations, the number of these locations, as measured on microarrays, can be quite considerable. On the other hand, a likelihood function for the binding sites based on a training set provides a statistical method of learning the pattern of the binding site and hence assesses the likelihood of any new binding site. Both methods have their own pitfalls: ChIP-on-chip suffers from low resolution due to DNA sonication, while likelihood functions are limited by the training set and methodologies used. In this work, I assess the linearity between the likelihood function and ChIP-on-chip signal peak as a function of binding site affinity and present a new method of combining these two approaches to consider predictions that are confirmed by both ChIP-on-chip signal cut-off and a likelihood function cut-off.

## 4.2. Introduction

As explained in the thesis introduction ChIP-on-chip (Aparicio et al., 2005, Ren et al., 2000) is a method that uses Chromatin Immunoprecipitation with microarray technology to detect the locations of the binding proteins on the DNA. ChIP-on-chip protocol as summarized in the introduction would generate a fluorescence image showing putative locations of the binding sites for a particular transcription factors.

The resulting fluorescence image is then quantified (after being normalized versus the control signal) representing the signal strength due to binding of the protein to DNA for every chip probe which is assigned to a start and end position on the DNA. These numbers are then statistically analyzed to detect the peaks that identify the DNA positions where the protein of interest is binding across the genome.

Peak detection algorithms and programs are underdeveloped, as the methods have been recently introduced. There have been some recent series of attempts to develop a suite for the analysis of ChIP-on-chip signal by Benoukraf, et. al (Benoukraf et al., 2009), CoCAS which is developed as a package in R building upon BioConductor (Gentleman et al., 2004). A previous notable work in peak detection is MPeak (Zheng et al., 2007), which recognizes peak shapes, then post processes the shape to identify the binding site. In addition, there is Ringo (Toedling et al., 2007) which is also a package based on BioConductor.

ChIP-on-chip protocol presents a direct way of detecting the binding of the protein of interest to the DNA, but the method poses resolution limitations as it relies mainly on DNA sonication

to produce fragment sizes, which are limited to 200 base pairs. The fragmented DNA is then hybridized to 500 to 1000 base unique DNA fluorescent fragments. This limitation renders the detection of the peaks the most challenging part of the protocol, as the signal will be generally characterized by wide base of the signal triangle, where a signal can be resulting from a binding event, which is within probe size upstream/downstream. For instance, a binding at base ***x*** on the DNA, which is hybridized to 500 bp fragments, can generate a signal in a probe, which is 500 bp upstream of ***x*** or 500 bp downstream of ***x.*** Another limitation would be the noise resulting from the low specificity binding events, which can be non functional sites as concluded by yong (Li et al., 2008). These limitations suggest that ChIP-on-chip signals need to be correctly corresponded with other tools to filter the binding events detected by the method.

In this work I have: First) tried to find a correlation between the signal strength and the predicted binding site statistical likelihood, with the hypothesis that a higher binding signal is a result of a high concentration of the protein of interest and hence indicates high affinity. Such affinity should correlate well with the binding site likelihood. In other words, higher predicted likelihood of a binding site should indicate higher signal in the chip arrays corresponding to such binding site. This correlation was applied for likelihood of each of the prediction methods explained before to test the superiority of any of these methods in correlating with the ChIP-on-chip signals. Second) provided a method to assign each putative binding site likelihood starting at position *i* with a ChIP-on-chip signal, then optimized a set of cut-offs for both the signal and the likelihood based on known binding sites to filter the predicted binding site by likelihood cut-off with a signal confirmation from a ChIP-on-chip experiment.

## 4.3. Methods

Generally the ChIP-on-chip fluorescence signal data were obtained from previously published work, so for LexA, I have used the work published by Wade et al. (Wade et al., 2005) and for the CRP binding site I have used the work published by Grainger et al. (Grainger et al., 2005).

### 4.3.1 Assignment of ChIP-on-chip signal to binding sites.

A ChIP-on-chip signal is assigned to every binding site, through an iterative greedy algorithm: 1) Given a set of putative binding site start positions *x*, 2) Consider a window of normalized probes including (*x* - probe size) to (*x* + probe size), 3) Assign the highest signal peak found within this window to the binding site (so this is the hypothesized binding site that generated this signal). This algorithm is applied over whole data set of normalized ChIP-on-chip signal probes.

The result is a pair wise assignment of binding site likelihood to signal strength within the window range. Apparently, this algorithm might not result in an optimum solution since the binding sites could overlap and two binding sites can be located within the window used, where both will be assigned to the same signal. This also is partly due to the deficiency of ChIP-on-chip method resolution as discussed before.

The whole genome of *E. coli* K12 MG1655 was scored against various likelihood methods for every position in the genome extended to the length of binding site. This score is then linked to

the signals using the above method and a matrix of ChIP-on-chip signal strength versus likelihood is obtained.

#### 4.3.1.1. Optimizing cut-off selection

A cut-off for both ChIP-on-chip signal and binding site likelihood is selected to achieve the optimum sensitivity and specificity for predicting binding sites for both of them. To assess the specificity and sensitivity of the method, I used the known binding sites in both LexA and CRP true positives. The method explained above is then applied over known binding sites to detect the highest/closest peak. An iteration is then applied over both signal cut-offs and likelihood one starting from the mean of the signal/likelihood distribution +1% area under distribution curve every iteration (i.e. 50 iterations).

The cut-offs are chosen such that the known binding sites detected above both cut-offs are considered as true positives (in the upper right quadrant in figure 4-1); the other known binding sites detected below both cut-offs are considered false negatives (in the lower left quadrant in figure 4-1); the unknown binding sites detected above both cut-offs are considered false positives; finally, the unknown binding sites detected below the cut-offs are considered true negatives. The cut-offs were optimized simultaneously for the best sensitivity and specificity to find thresholds that maximized their product.

### 4.3.2 Collective correlation between ChIP-on-chip signal and likelihood

The linear correlation between the signal data and the likelihood is facilitated using the signal analysis tool CoCAS (Benoukraf et al., 2009). CoCAS assigned a score for selected signal peak ranges, giving a range of probes with collective signal strength. This approach is summarizing the signal strength for every range of probes (given by a start and end position on the DNA), where it is believed that binding sites exist. To correlate such a collective model of signals, I have summed all the log likelihoods for putative binding site starting at every position in the range given as follows:

$$\rho(x, n) = \sum_{i=x}^{x+n} Likelihood_{bs}(i, i + l) \quad \textbf{(1)}$$

Where Likelihood is a likelihood function of any of the previously explained methods in chapter (3), x is the start position of the probe range given by CoCAS, n is the number of nucleotides involved in the range, $l$ is the length of the binding site.

## 4.4. Results

### 4.4.1 Combining likelihood with ChIP-on-chip signal improves prediction of known binding sites

Combining likelihood with ChIP-on-chip has enhanced the prediction of known binding sites. The idea behind this method is combining computational and biological prediction data to confirm the binding site prediction with an enhanced prediction power. As described in the methods, the cut-off for both methods (vertical and horizontal lines) are optimized to achieve the best specificity and sensitivity for detecting known binding sites (True positives, upper right quadrant).

The variance in the prediction accuracy given the method is very low (var = 2 %) similar to the variance noticed using the computational method alone as shown in Figure (4-2) and Table (4-1). This indicates the minor effect of ChIP-on-chip signal to either enhance or worsen a specific computational method.

The conservation of the binding site on the other hand seems to affect the outcome accuracy, as for LexA, the higher conservation in the binding site enhanced the specificity of the true binding site prediction, while for the promiscuity of the CRP binding site, the specificity and sensitivity is worsened by an average of 10% which emphasizes the effect of the promiscuity of the binding site on the ChIP-on-chip to generate a noisier binding profile.

Inspecting the performance of various methods carefully, I can notice that ULPB method provides a higher prediction power for LexA than the rest of the methods with 73% sensitivity and 99% specificity as shown in (Figure 4-1-B)

For CRP on the other side, I can notice the performance of EDNA is better (62 % sensitivity) in enhancing the sensitivity of the method than the rest of the methods (~ average 60% ) while the specificity of Dialign and ClustalW is better (88%). ULPB on the other hand is able to predict known binding sites with approximately equal sensitivity to UJP as shown in (Figure 4-3) and Table (4-1).

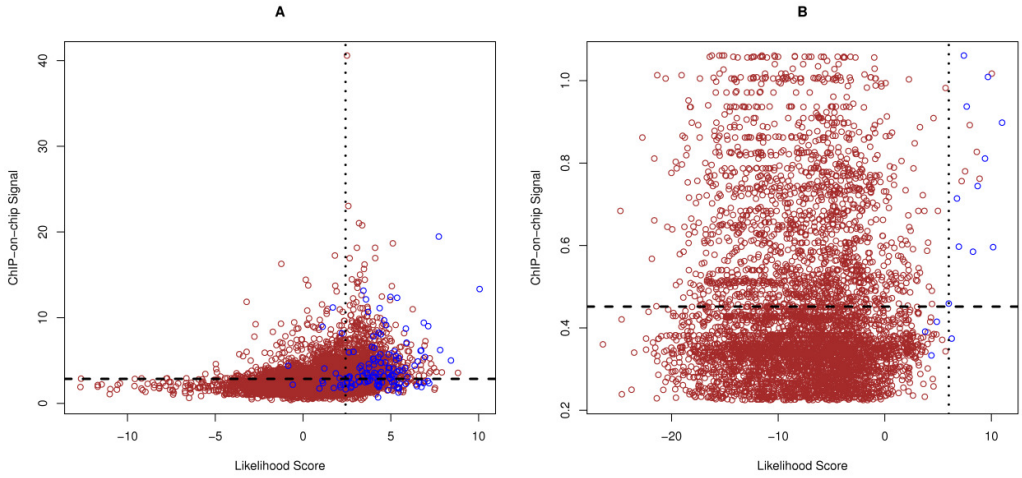


**Figure 4 - 1**: A) ChIP-on-chip analysis of CRP linked with the whole genome likelihood according to ULPB method. Blue dots shows probes corresponding to known binding sites, and other probes on the chip in brown. The horizontal line shows the optimal signal cut-off and the vertical line shows the optimal likelihood cut-off. B) ChIP-on-chip analysis of LexA linked with the whole genome; details as in (A).

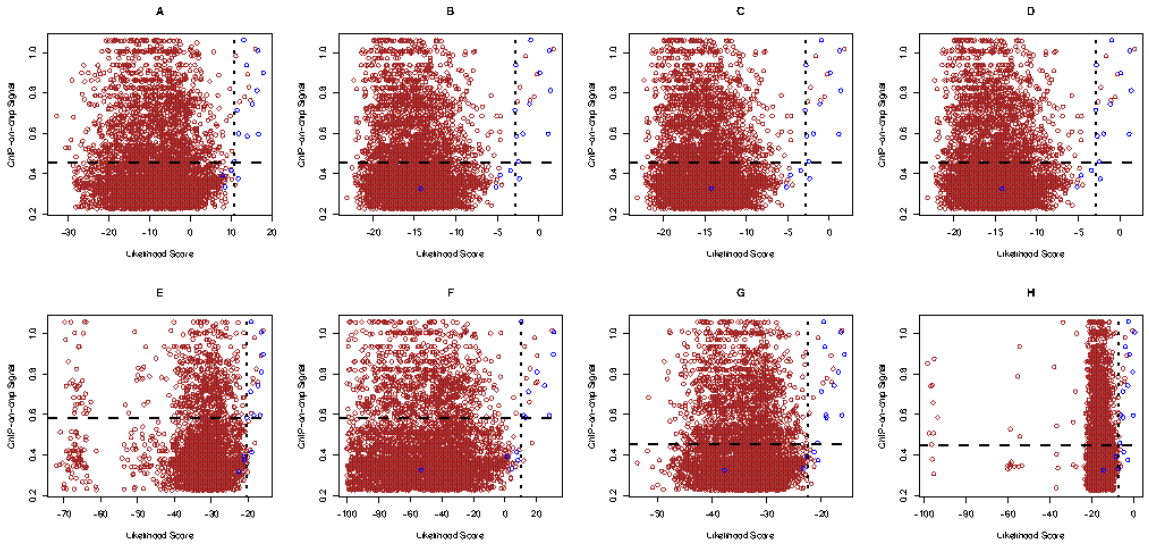


**Figure 4 - 2**: A figure showing the ChIP-on-chip analysis of LexA linked with the whole genome showing probes corresponding to known binding sites as blue dots and other probes on the chip in brown. The horizontal line shows the optimal signal cut-off and the vertical line shows the optimal likelihood cut-off. This is shown for A) ULPB, B) Dialign, C) EDNA, D) SDNMSA, E) PSWM, F) UJP, G) Ungapped, H)ClustalW.

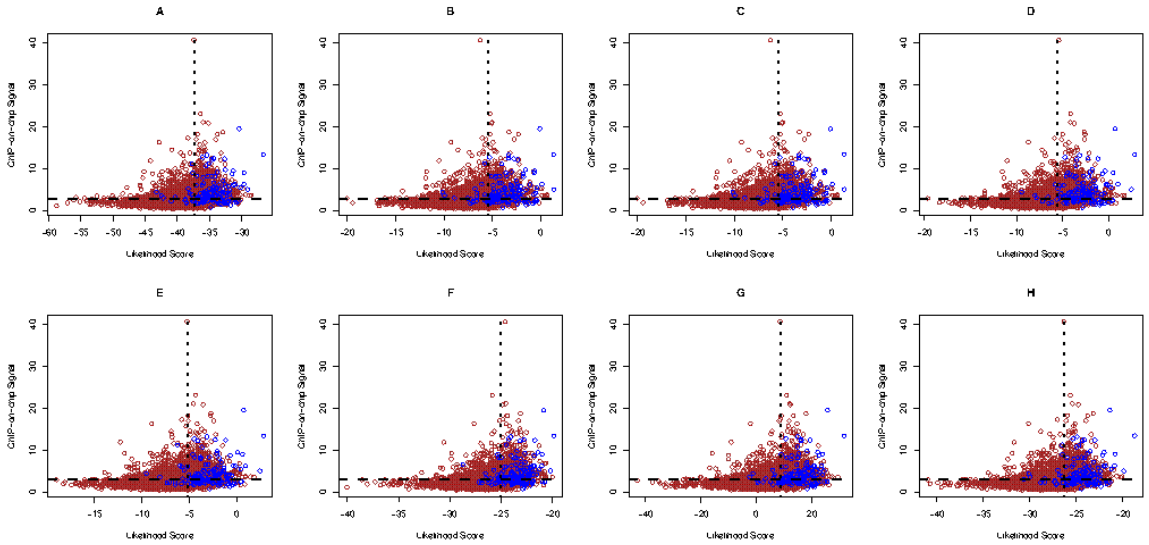


**Figure 4 - 3**: A figure showing the ChIP-on-chip analysis of CRP linked with the whole genome showing probes corresponding to known binding sites as blue dots and other probes on the chip in brown. The horizontal line shows the optimal signal cut-off and the vertical line shows the optimal likelihood cut-off. This is shown for A) ULPB, B) ClustalW, C) CDialign, D) EDNA, E) SDNMSA, F) PSWM, G) UJP, H) ungapped.

| Binding Site | CRP | | LexA | |
|---|---|---|---|---|
| | *Sensitivity (%)* | *Specificity (%)* | *Sensitivity (%)* | *Specificity (%)* |
| PSWM | 58 | 87 | 69 | 99 |
| ClustalW | 60 | 89 | 68 | 99 |
| UJP | 60 | 87 | 64 | 99 |
| Ungapped model | 58 | 87 | 68 | 99 |
| ULPB | 58 | 87 | 73 | 99 |
| Dialign | 60 | 89 | 68 | 99 |
| SDNMSA | 60 | 88 | 68 | 99 |
| EDNA | 62 | 85 | 68 | 99 |

**Table 4 - 1**: Sensitivity/Specificity analysis of CRP and LexA linked with the ChIP-on-chip signal

### 4.4.2 ChIP-on-chip signal regresses linearly with likelihood

The linear collective regression between the ChIP-on-chip signals has been found to fit well with collective likelihoods as shown in (Figure 4- 4) and (Table 4-2). It has been found that likelihood generated from various methods did not provide any significant differences in this regression. LexA has been found to have a higher R-Square of the linear regression than CRP. In CRP ULPB has lower R-Square than other methods of 0.61, which have equal R-Square of 0.65.

The LexA linear regression has shown only 5 points of false in-silico positive predictions. Those outliers are mainly regions that has low signal in the ChIP-on-chip although coming up with high likelihood values. On the other hand, CRP suffers from a larger number of outliers, and the outliers in this case are mainly under estimations from the in-silico predictions where a higher ChIP-on-chip signal is detected while a lower (or fixed) likelihood is predicted. Those regions mainly reflect the deficiency in the in-silico prediction to capture the actual binding specificity using a statistical method. Finally, we can notice the worse behaviour of the first order HMM confirming the argument in the previous chapter that the MSA is disrupting the actual binding site interdependencies.

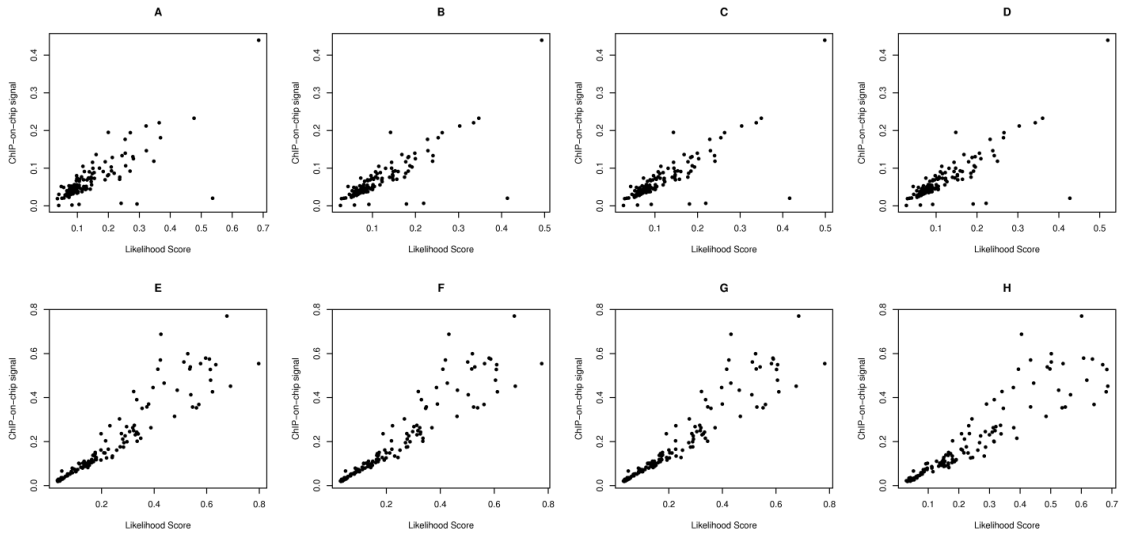


**Figure 4 - 4**: Correlation between the ChIP-on-chip signal and the various likelihood scoring functions in the order of ULPB, PSWM, ungapped, 1st order HMM, for both CRP and LexA. Figures A-D is for LexA, while E-H is for CRP.

| Binding Site | PSWM | First order HMM | Ungapped model | ULPB |
|---|---|---|---|---|
| CRP | 0.65 | 0.65 | 0.65 | 0.61 |
| LexA | 0.87 | 0.83 | 0.87 | 0.87 |

**Table 4 - 2**: R-Squared evaluation of the linear regression model for binding sites linked with the signal from CoCAS

## 4.5. Discussion

### 4.5.1 Linear regression between ChIP-on-chip signal and binding site likelihood cannot be established for every binding site.

The combination of the statistical likelihood with ChIP-on-chip signal of binding sites is expected to filter out from both methods the false positives, false negatives and increase the number of true positives and true negatives. An initial hypothesis is that the ChIP-on-chip signal must be correlated with the binding site likelihood, through affinity of the binding site, such that higher affinity should explain the more likely binding sites and as well should explain the ChIP-on-chip signal strength, since high affinity of the binding sites would increase the copy number of protein of interest localization in this site and hence increase the fluorescence signal intensity. Albeit such an obvious relation, the linear correlation between the likelihood and the signal cannot be established for every binding site, and obviously, this is due to the resolution of the method, which has been the problem of this method from day 1. The low resolution of the method allows the same signal to be a result of multiple binding sites and hence the intensity of the chip probe can be rooted back to multiple binding sites within the window of this probe as explained in the algorithm.

On the other hand, this linear regression can be better established over the multiple probes and multiple binding sites in the same region as explained in the methods. This regression have shown high R-Square for LexA which is a highly conserved binding site, but lower for promiscuous CRP binding site.

The lower R-Squared for the CRP binding site can be attributed to the lower prediction power of the likelihood function that is not efficient enough to learn the global promiscuous pattern of the binding site, which will affect the linearity between the likelihood function and the binding site affinity. As a result, binding sites likelihood will be linearly correlated with binding sites affinity if highly conserved but with the loss of such conservation, likelihood will have to be highly efficient to learn the promiscuous global pattern of the binding site rather than leading to the average distributions of the bases in each position.

### 4.5.2 ChIP-on-chip and likelihood functions provides better accuracy for conserved binding sites

Prediction of binding sites have been always a problem, because of the promiscuous nature of the binding sites exhibiting various patterns of nucleotides which will all lead to binding but with variant affinity. ChIP-on-chip provides a biologically verified method of finding the sites where the protein of interest binds on the DNA molecule. Although this method provides a list of prospective locations, the number of these locations can be enormous. Yong (Li et al., 2008) have proved that some of the sites discovered can be non functional sites, others have postulated that the protein of interest can be linked to the chromatin structures as in CRP (Grainger et al., 2005), which are all perfectly logical explanations, except for the fact that it renders the discovery of the binding sites quite hard even with using such method.

On the other hand postulating a likelihood function for the binding sites based on a training set provides statistical method of learning the pattern of the binding site and hence assesses the

likelihood of any new binding site. The problem with using such a method is that because it is based on learning, it suffers from information generalization. Such a problem can lead the likelihood to be as good as its training set and hence can be easily stuck in local maxima rather than approximating the global pattern of the nucleotides. This problem will always drive more researchers to think of better methods to assess the likelihood of the binding site.

A simple method used in this work, is to filter those sites predicted by the likelihood functions with those found by the ChIP-on-chip and vice versa. Assessing this method for two binding sites I have found that for LexA, ULPB was able to predict more biologically verified binding sites than any of the other methods. For CRP, ULPB was as good as UJP and 1$^{st}$ order HMM in finding the true positives, while PSWM has proved to be the best one in discovering the true negatives.

The reason for such low prediction power is that the optimization algorithms for the cut-offs used is a greedy algorithm, in other words, it tries to find the highest ChIP-on-chip signal that maximizes the fitness function (Sensitivity * Specificity). Apparently, many of the binding sites known will have low affinity leading to low signal peaks, which will be then filtered out by this algorithm or any other algorithm looking for the binding sites in ChIP-on-chip peak lists. This can be noticed in highly conserved LexA which have a much higher fitness than CRP, this conservation goes well with the greedy algorithms, since the variance of the peak signals across the true positives will be lower than that for a promiscuous binding site like CRP which is expected to have higher variance of peaks across the known binding sites.

Such a work suggests that ChIP-on-chip fits well the discovery of conserved binding sites rather non conserved one, and suggests that likelihood function needs to be more robust in identifying global binding site patterns.

PART II

# GENE REGULATORY NETWORKS MODELLING

# Chapter 5

# COMPOUND ORIENTED MODELLING

## 5.1. Abstract

Biochemical network modelling includes the challenging research of transforming a biological problem into a computational problem for *in silico* simulation. The challenges arise mainly from the difficulties faced by the researchers in modelling the networks and getting the intended answers for simulation questions. The modelling act involves a description of the network and its topology and then simulating it *in silico*. The current modelling grammars that describe the networks suffer from many defects, including lack of reusability, generalizations and combinatorial expansion of reactant states that can render them a barrier for *in silico* modelling. In this work I present a new a modelling paradigm, compound oriented modelling which shifts the modelling process from being a reaction focused to being compound focused. This modelling paradigm is presented as a new grammar that helps address the defects in the current languages. A gene regulation network is presented as an extension to the basic grammar, and a case study of MelR gene regulation is presented at the end of the chapter.

## 5.2. Introduction

Biochemical networks are comprised of biochemical compounds and reactions involved together in complex biological processes formed by much simpler processes where the output of one reaction can be the input to another. Modelling of these networks is one of the most important research lines in computational biology. Its importance lies in the fact that, first, computational modelling of a biochemical network allows deeper understanding of it, and second, simulating the network *in silico* can generate hypotheses to be tested either *in vivo* or *in vitro*.

The techniques used to simulate the dynamics of the network are either deterministic or stochastic. It has been shown that stochastic simulations provide important dynamic effects to the simulation of many biochemical networks, e.g. as phage λ switch decision between lysis and lysogeny (Mc Grath and Sinderen, 2007). A major research direction for stochastic modelling of biochemical networks is the use of Process calculus that has been used originally in modelling mobile communication, as it shows an advance over lambda calculus in concurrent processing by message passing. The analogy between mobile communications and biological pathways has drawn the attention of computational biologists to formulate biological processes with process calculi (Sangiorgi and Walker, 2001). Process calculi have proved to be useful in modelling molecular interactions (Priami et al., 2001, A. Phillips and Cardelli, 2004, Calder et al., 2006, Priami, 2005) and specifically biological signalling pathways and gene regulatory networks, albeit being very complicated in syntax. Multiple interfaces have been presented to harness the power of the Process calculus and present a user friendly modelling

language (Kahramanoğullari, 2009) in an attempt to tackle such complexity, and enable biological researchers to benefit from its potential.

In this chapter, I focus on the modelling part by introducing a new modelling grammar that enhances the modelling process allowing the modeller to capture the fine details of the simulated subject in a minimum error prone process.

### 5.2.1 Current problems in modelling languages

Model description provides a route for transforming a biologically motivated problem into a computer problem that can be simulated in a biologically meaningful manner using any of the previously described approaches. The common use of informal models in biological literature to describe the network has proved to be confusing and ambiguous in complex systems and may hide information that is scientifically important for computer simulation (Guerriero et al., 2007). The formal modelling languages on the other hand guarantee the correct transformation process of cellular descriptions to computational descriptions but lack many important features that can help them being an enabler rather than a barrier indeed.

#### 5.2.1.1. Computer intuitive languages

Current languages are mostly computer intuitive rather than being biologically intuitive (Kholodenko, 2006) requiring a good knowledge of computer systems. Some researchers have tackled this problem simplifying the grammar of the language into a narrative paradigm (Mc Grath and Sinderen, 2007).

#### 5.2.1.2. Extension

Modelling languages in general tends to abstract the modelling units available for the modeller to use, for example in SBML, the main modelling units are reaction, reactants and compartments. This abstraction defines the framework for the programmer to think about his model, so for example to model a gene transcription process, the modeller would abstract all the details in terms of reactions and reactants. So for instance accommodating complex scenarios in SBML can be a lengthy error prone process, since complex interactions can be involved. Hence, the modelling units defined by the modelling language would determine the capacity of the language to accommodate various scenarios.

This problem can be easily tackled by giving the modeller the utility of defining their own modelling units using extension and the modeller can then increase the specificity and complexity of the modelling unit allowing increased capacity of the model to accommodate complex scenarios while not losing abstraction.

#### 5.2.1.3. Combinatorial expansion of the chemical reactions

Combinatorial expansion of the chemical reactions involved in a certain operation tends to manifest itself in current languages when used to comprehensively model reactant states involved. SBML for instance, although being standard and abstract enough to model any biochemical network type, it is producing a model description whose complexity increases exponentially with the number of reactants introduced to the network. This complexity originates from the fact that it does not assume implicit reactions, which can be assumed if the

language targeted a specific type of networks. For example, to model all the possible transcription reaction rates of one gene with n binding sites, the modeller will have to specify all the combinations of the m transcription factors binding to n binding sites and their effect on the transcription rate, while in essence only a handful of cases change the transcription rate and the other states can be given one value for their transcription rate.

#### 5.2.1.4. Reusability

Topological structures in biochemical networks are repetitive and can be generalized to simpler units. This feature of biochemical networks allows for high reusability in the topology description. For example the complex structure of gene regulatory networks is composed of simpler repetitive motifs (Alon, 2006), such as feed forward loops and bi fans. These motifs can be regarded as reusable function units given the variable reactants connecting them. This reusability allows the modellers to model these motifs once, and then use them as a library, which can also be shared between other modellers.

#### 5.2.1.5. Coupling between language and simulation paradigm

Coupling between the modelling language and the simulation paradigm in a single platform, such as Cell Illustrator (Matsuno, 2002), JigCell (Vass et al., 2004), METATOOL (Pfeiffer et al., 1999), CellDesigner (A. Funahashi et al., 2003), COPASI (Hoops et al., 2006), E-Cell (Tomita et al., 1999), BIOCHAM (Calzone et al., 2006) or JDesigner (Sauro et al., 2003), NIRest (Lauria et al., 2009) or Arcadia (Villeger et al., 2010), without portability to the others, ties the modeller to both the advantages and disadvantages of this platform. Although this

problem has been lately tackled to an extent with most of these platforms supporting SBML export/import, the process of exporting SBML from one tool and importing it into another tool is far from being straightforward as it suffers from tool/SBML personalization of the way each tool transforms their model to SBML modelling units while exporting it.

### 5.2.2 Object oriented inspired modelling

The modelling language problems described above has led various computer scientists to adopt an object oriented inspired paradigm to solve such problems, as in the work of (Webb and White, 2006) describing an object oriented approach to model the cell using UML as objects and agents which present a software engineering approach to model the cell in a top down systematic series of steps. This approach mainly suffers from being computationally intuitive rather being biologically intuitive. A functional inspired approach was adopted by Sauro (Sauro, 2006) which presents cellular modules as functions with inputs which can be reused, and hence solves the reusability issue. This language mainly suffers from being none biologically intuitive language and does not solve the extension problem nor does it solve the combinatorial expansion problem.

### 5.2.3 Narrative compound oriented grammar

Prompted by this gap, and inspired by existing languages, I introduce in this chapter a grammar for biochemical networks that has a hybrid style of both a narrative and a functional nature (Barendregt, 1984). This grammar presents solutions for the following problems in current grammars.

#### 5.2.3.1. Biologically intuitive grammar

The narrative style of the presented language bridges the gap between a formal computer language and a biologically intuitive language. The biological intuition is reflected in the grammar syntax, which reflects a biological scenario rather than computer steps. This is mainly reflected in the narrative conditional style that reflects an event based approach. So for example, the modeller can specify what should happen if a transcription factor binds to a certain gene, he can then specify a dependent binding event or actually specify the transcription event of the gene.

#### 5.2.3.2. Functional nature of the language is inherently reusable

The functional nature of the presented language originates from the fact that the reaction can be regarded as template reactions as in functions in programming languages, which brings all the powers and features of the functional programming to the biochemical networks. The reaction writing in this case is split into two steps; first step is writing the reaction, which can be a cooperative transcription of two factors engaged in cooperative binding. In this case, the modeller will assume two transcription factors and then write an event for the binding of these two transcription factors and how much they will affect the transcription rate of the reaction. Second step will be using this reaction, since this reaction assume any gene and any two transcription factors, then the modeller can use it with any gene and two transcription factors that fits his current model, hence the modeller can reuse his reaction over and over again without having to write it every time. This functional reusable feature have been always used in

programming languages, in this language I try to harness this feature to solve the reusability issue raised by other languages.

#### 5.2.3.3. Compound oriented modelling provides a mean for extension

Functional programming alone cannot present a solution to extension hence in this chapter I introduce the notion of chemical compound oriented programming, where the chemical compound provides a complete unit describing the properties of the chemical compound and the reactions it is involved in. The concept of the chemical compound is introduced into this language to shift the modelling orientation from reactions to reactants, allowing for, 1) separation of logic between the chemical compounds, grouping the reactions as a property of the chemical compounds. 2) extension of the chemical compounds to any modelling unit. This basic chemical compound type can be extended to any domain, so the chemical compound can be a Protein, DNA, Gene etc. and the reactions will be the associated set of reactions for this chemical compound. For example, a protein as a chemical compound can be associated with a set of reactions specific to proteins like polymerization.

#### 5.2.3.4. Default cases minimizes the combinatorial expansion effect

The narrative control logic of the language allows for default cases where the reaction rates are not affected, solving the problem of the combinatorial expansion of the reactants relationship. This feature minimizes the effort the modeller has to put to model all the cases. The language in this case normally generates all the reactions with the specified default reaction rates. For example, in gene transcription, the transcription rates of gene with bound transcription factors to binding sites can be described using all occupancy states, where only one transcription factor is bound, or two or many, so unless the modeller specify a specific case for the transcription rates which can lead to activation or repression, all the of the occupancy states of the gene will be automatically generated for the modeller with the default rate.

#### 5.2.3.5. Decoupling the language from its output allows for multiple simulation paradigms

The language has a high level specification that decouples it from its output, but since some properties of either the chemical compound or the reactions are entirely output specific descriptions, I have included the possibility of adding annotations to both the chemical compounds and their reactions to keep the language as simple as possible. The annotation will be processed by the translator controlling its translated output as shown in the next section. The translator of the language is built with a layered architecture decoupling the actual parsing of the language from its mapping allowing the programmer to implement translation functions that map the reactions to other modelling languages. The current given translator translates the

model to SBML to allow initially for model portability, equally, a translator for matlab code or any other language could be implemented.

### 5.2.4 Gene regulation networks

There have been many efforts to model the biochemical networks in the biological literature, many of which focused on signalling pathways and only a few for gene regulation (Schlitt and Brazma, 2005, Smolen et al., 2000, Wahde and Hertz, 2001). In this chapter, I introduce a basic library for gene regulation networks, which can then be extended by modellers to suit their needs. I also introduced a model of *MelR* (Webster et al., 1988) gene regulation as a case study to show the use of the language with a real example.

## 5.3. Language Grammar:

### 5.3.1 Data Types:

**Compound:** This is the basic reactant type of the model. The compound can be a simple element or a composite element. A simple element is given a name. A composite element on the other hand can contain other compounds, reactions or constants. The composite elements can be parameterized, so for example a gene can be parameterized by its activators and repressors. The compound elements can be referenced using "***of***" keyword as in the example below. Should the composite compounds have variable compositions then the modeller can parameterize the compound with other compounds.

**Constant:** This is the data types used mainly in the mathematical functions or it can be used as the modeller needs.

***Example***

```
#BindingSite is a chemical compound with reaction Binding defined in the gene regulation.
Gene g {
        # describing gene compound termed g
        Constant c is 0.04 # defining a constant with value 0.04
        BindingSite bs1# defining a bindingsite bs1
        BindingSite bs2 # defining a bindingsite bs
}
```

### 5.3.2 Functions:

The functions in this language are descriptive functions. Working with functions requires first defining it using the word "**define**" followed by any given name then using this defined function by just writing its given name and the parameter values if any. Parameters are considered as placeholders that will determine the behaviour of the function. As for the reaction, the parameters provide a way to define preconditions that must exist for the reaction to occur. The functions also being referentially transparent allow for high order definitions (i.e. the function can be used as its result). The functions of this language are one of three types:

**Reaction:** this type of function is a property of the compounds and it has to be used on a certain compound reacting with other compounds.

Reactions can be parameterized by both compounds and constants. Reaction can be simple defining only what is reacting with the compound or it can be composite reaction defining a process of reactions until a final product is obtained. Results of the reactions are always regarded as compounds. Their aim is to describe the reactions among the reactants. They can be assigned many properties as the on rates (rate constant) and the off rates (stability constant) of the reactions and all the reactants. One reaction can be used as an input to another reaction since they are referentially transparent.

**Lambda***:* This is the type of function that can be parameterized by only constants.

Their aim is to model any mathematical operation while controlling the dynamics governing your model. For example Michaelis-Menten expressions.

**Network***:* Network type is a property of the model and the starting point of the model.

The model can contain as many networks as possible, which can contain each other. The un-included network is considered the main one. You can only have one main network in the model.

***Example***

```
 # describing gene compound termed template_gene and parameterize it with transcription factor that is needed for its transcription
Gene template_gene given (TranscriptionFactor TF1) {
        Constant c is 0.04 # defining a constant with value 0.04
        BindingSite bs1# defining a bindingsite bs1
        BindingSite bs2 # defining a bindingsite bs2
        Binding of bs1 using (TF1) #using binding reaction of compound bs1
}
#Define the network of your model
Network {
        Protein P1 #define Protein P1 to be used as a transcription factor
        template_gene using (P1) #use defined gene passing the protein as a parameter
}
```

### 5.3.3 Extensions:

*Compound Type:* The compound types are basically extensions of the language basic type Compound. The extensions in the language are implicitly determined if the modeller defined a certain compound and then starts defining its details as in the above samples of gene type. Some extensions are already defined for the modeller for gene regulation in a form of library,

for example, the gene type that implicitly has transcription, loop and translation_transcription reaction. All the reactions and lambda functions defined in a type can be simply used or overridden by the modeller by writing its name and describing new behaviour. Casting of the types is not offered since this is an implicit feature.

### 5.3.4 Control Structures:

The language contains only conditional control statements, which allow the modeller to model events conditionally.

***Statement**:* if .. then .. else/elseif …. then… endif

***Expressions**:* the expressions used in the condition statement are simply Boolean expression. There are two types of Boolean expressions:

***Reaction type**:*

- is (not) present: tests if the chemical compound is present at the moment
- is (not) boundto: tests the binding state of the compound to other compounds
- is (not) typeof: tests the type of the compound (for example is the transcription factor type of activator, repressor).

***Mathematical type**:*

- Not: negation of a certain condition
- Equals: tests the inequality of two expressions
- Greater than: tests if one expression is greater than another
- Less than: tests if one expression is less than another

***Example***

```
# describing new gene type called mutual and defining Mutual_ Transcription reaction, which
is a composite reaction
Gene mutual {
        BindingSite bs1 and bs2
        Define Transcription_Mutual given (Activator a, Activator b) {
                # describing the control logic as a sequence of events
                If bs1 is boundto a then {
                        Binding of bs2 using (b)
                        Transcription # using Transcription function already defined for type
                        gene
                }
                if bs2 is boundto b then{
                        Binding of bs1 using (a)
                        Transcription
        }
}
}
```

This example shows a definition of gene with a mutual transcription where the transcription rate is changed by the order of binding. The highlighted line shows the binding reaction of bs2 only if bs1 is bound to a.

### 5.3.5 Annotations:

The annotations are properties of either the chemical compound or its reactions that should be translated in the output model, annotations are used in the language with a prefix "**with ()**" and key value pair between those parenthesizes as shown in the sample below. The annotations used should be configured in another XML file, which maps the annotation name and the name in the output. This file is translator specific file, and this simple mapping behaviour in our translator can be overridden by advanced programmers with complex mapping functions. The basic annotations in the language are reaction rate for the reactions and concentrations for the chemical compounds. In the sample below, I have built on the previous example to add the reaction rate information to the control flow showing that the sequence of binding affects the transcription.

***Example***

```
Gene mutual {
BindingSite bs1 and bs2
Define Transcription_mutual given (Activator a, Activator b) {
        If bs1 is bound to a then {
                Binding of bs2 using (b) with (rate 0.01)
                Transcription with (rate 0.5) #adding annotation for the transcription rate
        }
        if bs2 is bound to b then{
                Binding of bs1 using (a) with (rate 0.02)
                Transcription using (RNAp) with (rate 0.1)}}
```

## 5.4. Gene Regulation Extensions:

In order to model gene regulation I do not need to change the grammar but rather use the extension feature of the language, and the compound types, as well as include the new types as a library. All the extensions are basically adding more specificity to the type of reactants reacting return type and default reaction rate while adding other reaction functions. All the below types the modeller can just use them as a basic function or can simply define more compounds inside it and more functions within .

**DNA***:* this is the normal DNA molecule.

**Gene***:* this is a type of DNA molecule and includes three extra functions; *transcription*, *transcription_translation* and *loop* returning gene

**mRNA***:* this is the messenger RNA molecule and has two functions; *degradation and translation*.

**RNAp***:* this is the RNA polymerase molecule and has one default reaction; *promotion*.

**Protein***:* this is the protein molecule and has one default functions *polymerization.*

**TranscriptionFactor***:* this is actually an extension of the *protein*, and has an extra function bind that can bind to a *BindingSite* type.

**Promoter***:* this is actually a DNA molecule and is part of the gene molecule.

**BindingSite***:* this is another DNA molecule and is also part of the gene molecule and has a default reaction function *bind*, which can react with any *TranscriptionFactor*.

**Inducer/Activator***:* this is a type of *TranscriptionFactor* and includes an extra function *Activation.*

**Repressor***:* this is a type of *TranscriptionFactor* and includes an extra function *Repression.*

**Dual***:* this is a type of *TranscriptionFactor* with no extra function and the use of this type will require the modeller to include either an *activation* or *repression* function.

**Cofactor***:* this is a compound, which includes extra function bind that can bind to *RNAp* molecule.

## 5.5. Language Features:

### A. Reusability

The reusability is defined in language through the possibility of reusing the already defined compounds and functions by using the word "*use"* followed by a model file. Any function or compound already defined in this model file can then be used in the current model. This will allow modellers to build on the past models if any resemblance is found.

### B. Model Checking:

The model checker implemented in the language is the most basic one, which checks the existence of contradicting reactions in terms of reaction rates.

### C. Parametric Polymorphism:

The parametric polymorphism in this language is defined to allow the modeller to define the same reaction name as many times but with different parameters.

### D. Generalization:

The abstraction of the compound type in the language and the ability to specialize from this general type to a more specific type is an important feature of the language. This feature allows

for extending the language to specific domains as the gene regulation without affecting the basic concepts.

## E. Ease of use:

The most important requirement of this language is to suit the biologists and enable a seamless bridge between the biology model and the computer model. The ease use feature is experienced through:

- **Distributive**: The language supports distributive law for Boolean logic so if the same Boolean test applies to multiple compounds then the modeller can use the distributive law.

  *(A ® B ® C) © D = A © D ® B © D ® C © D*

  ***Example:***

  *If a and B or C isboundto D => if a isboundto D and B isboundto D or C isboundto D*

- **Typing**: Name without a type is assumed to be basic chemical compound type unless it exists in the same statement with other type except for reaction definition.

  ***Example:***

  *Gene A, B => B is Gene*

  *C => C is chemical compound*

  *Define reaction given (Gene A, B) => B is a chemical compound not gene*

- **Naming**: Chemical compound types used without a given name, will be given a name by the translator derived from the type name.

- **Default Control**:
    - Control logic uncovered cases will be defaulted with other possible reactions given default reaction rate unless suppressed by modeller.
    - Reactions with multiple reactants, without control logic specifying reaction order, are assumed to be a reaction sequence based on the order of the reactants given.
- **Parameterization**: Reactions used without given parameters are assumed parameters from the types defined.
- **Default properties**: the concentration of the reactants is given the default concentration if not specified

## 5.6. Language Translator

The language translator is designed using the pipeline architecture as shown in Figure 5 - 1 and is built using Java programming language. The pipeline architecture breaks processing into a sequence of simpler transformations, each processing unit, termed “filter”, processes the model then passes it to the next filter through a pipe, which transforms the previous filter output to the next filter input. This architecture allows for highly decoupled configuration between the translator processing units, which allows the programmers to add, remove or replace layers of the translator. This also allows the translator to integrate with other packages, for example for stochastic simulation of the output of the model.

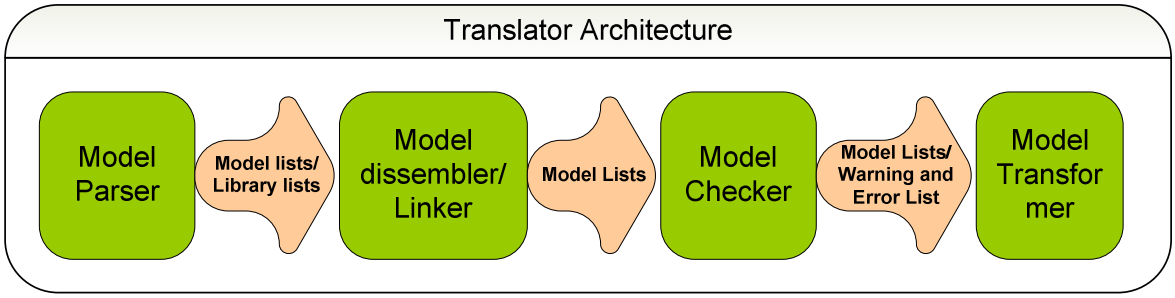


**Figure 5 - 1**: This is a simple sketch of the translator architecture, which shows a pipeline pattern.

The first filter of the model is the lex parser that parses both model and libraries into java objects of reactants and compounds. The second filter completes the model through two steps; 1) links the used libraries to the model and replace reusable parts with the actual model elements, 2) dissembles all the control structures into reactions and completes the missing model information (if possible) using the default behaviour. The third filter checks the model

integrity, in case there are any contradicting reactions before forwarding it to the next filter. The last filter is the model formatter, which is a set of interfaces that the programmer can implement to transform the model into any format, the current example implementation is SBML.

## 5.7. Case Study: *Escherichia coli MelR* gene regulation

In this section, I describe a case study for describing *MelR* gene regulation in *Escherichia coli*. The *melR* gene encodes the protein MelR which is a member of the family AraC/XylS (Gallegos et al., 1997).

The following model describes the gene MelR with its binding sites and its transcription control logic. Following is the basic suggested steps for defining a model for the gene regulation.

**Model Definition**: The model starts with the use of the library GeneRegulation.model that contains all the extensions of the gene regulations. The next statement is the start of the model giving it a name **MelR.**

**Gene Definition:** Typically, the next statements in the model are to define a gene type defining all the binding sites of this gene and define the transcription logic as detailed below. The *melR* gene has 7 binding sites, which defined as shown by the type **BindingSite**.

**Transcription Function:** The transcription function defined using the word "**Define**" denoting the definition of a new reaction. This reaction is parameterized by the transcription factors needed for the transcription to occur. These parameters are preconditions that should be supplied later when using this reaction. Annotations used with the definition of this transcription reaction denote the default rate and stability for all the reactions listed in this reaction if not specified.

**Control Logic:** The control logic defines the reaction sequence and the resulting transcription of this gene, which is biologically intuitive. The first statement for example specifies that if

CRP-CAMP1 binding site is bound to CRPDimer then the gene is activated, where if in addition MelrbsR and Melrbs2 binding sites are bound to MelR then you get repression. The last set of reactions specifies the rates of the unbinding reactions of these binding sites.

**Network:** The Network is the starting point for the translator. It prepares all the transcription factors needed for the melR gene transcription function, and then uses the gene and transcription function. This last statement in the main function will force the translator to translate the transcription reaction with all included reactions into the final output of the model. Notice the definition of MelR TranscriptionFactor with initial concentration then stating later that it is the result of the transcription reaction of the melR gene, which would show as an auto regulation in the gene simulation

An abbreviated version of the SBML model resulting from this model is included in Appendix II showing the obvious value of the model simplicity language when transformed into complex and lengthy SBML model and it also shows the combinatorial expansion describing default transcription reactions from various occupations of binding sites using the default transcription rate.

```
use GeneRegulation.model

Model MelR {
        Gene melR {
                BindingSite MelrbsR, CRP-CAMP1, CRP-CAMP2, Melrbs2, Melrbs2',
                Melrbs1, Melrbs1'
                Define Transcription_Melr given (TranscriptionFactor MelR,
                TranscriptionFactor CRPDimer, TranscriptionFactor Melr-Melibiose) with
                (rate 0.1 and stability 0) {
                if CRP-CAMP1 is boundto CRPDimer then
                        Transcription with (rate 1)
                else if CRP-CAMP1 is boundto CRPDimer and MelrbsR and Melrbs2 isboundto
                MelR then
                        Transcription with (rate 0.001)
                 End if
                Binding of Melrbs1 given (Melr) with (stability 0.2)
                Binding of Melrbs2 given (Melr) with (stability 0.2)
                Binding of Melrbs1' given (Melr) with (stability 0.3)
                Binding of MelrbsR given (Melr) with (stability 0.1)
                Binding of CRP-CAMP2 given (CRPDimer) with (stability 0.5)
                Binding of CRP-CAMP1 given (CRPDimer) with (stability 1)
                Binding of Melrbs1 given (Melr-Melibiose) with (stability 0.8)
                Binding of Melrbs2 given (Melr-Melibiose) with (stability 0.8)
                Binding of Melrbs1' given (Melr-Melibiose) with (stability 0.9)
                Binding of Melrbs2' given (Melr-Melibiose) with (stability 0.9)
        }
}
Network Melnet with (rate 0.01) {
                Melibiose with (initial_concentration 1)
                CRPDimer is reaction of CRP given (CRP) with (rate 0.1, stability 0.5)
                TranscriptionFactor MelrTF with (initial_concentration 5)
                TranscriptionFactor Melr-Melibiose is reaction of MelrTF given (Melibiose)
                with(rate 0.5, stability 0.1)
                MelrTF is Transcription_Melr of MelR given (MelrTF, CRPDimer,
                MelrMelibiose)
        }
}
```

## 5.8. Conclusion

In this work I have presented a new modelling paradigm based on compounds shifting the modelling orientation from reaction based to compound based helping the modellers to separate the reactions of each compound as property to that compound. I have also presented a grammar that helps solving current problems in modelling languages as reusability that allow the modellers to build libraries of common topologies for example and combinatorial expansion using control logic with default cases that help the modeller to focus on the special cases only. This language is extended as a library to model the gene regulation network where a case study of *MelR* gene regulation is presented. I have also presented a translator that is based on a pipe and filter architecture, which allows for plugging other modules that translates the reaction sets to any other modelling language. The current supported modelling language is SBML, which allows for portability of the model to a wide range of simulation environments. I anticipate that the paradigm shift from reactions to reactant oriented description of biochemical networks might be as important as the analogues shift from functional to object oriented programming languages for software development

# Chapter 6

# CONCLUSION

Gene regulatory networks have been the focus of biological research for decades since the discovery of the first network (Jacob and Monod, 1961). The objective of this thesis is to describe novel techniques to, first, to enhance the discovery of the networks through binding sites prediction and second, to enhance the modelling for *in silico* simulations.

In the first part of the thesis, to enhance the discovery of gene networks through binding sites prediction, I have introduced a number of novel methods that consider the dependence between neighbouring nucleotides as extensions of the conventional independent methods. The ungapped likelihood under positional background (ULPB) considers a 16X16 weight matrix where the values are conditional probabilities of nucleotides rather than a simple 4X4 position specific base probability. In contrast, the gapped $1^{st}$ order hidden Markov model considers a state for every nucleotide rather than a simple match state for all nucleotides considered in the gapped $0^{th}$ order Hidden Markov Model. By testing the prediction power of these models on 22 global binding sites in *E. coli,* I demonstrated the superiority of the ungapped method (Chapter 2). The superiority of the ungapped model over the gapped model was attributed to the deficiency in the alignment, which aligns the binding sites using an evolutionary hypothesis for binding sites in the same species.

This deficiency has led us to hypothesize a thermodynamic approach for binding site alignment based on stacking interactions between neighbouring bases. I have presented two new methods to align binding sites to test this hypothesis. Both methods (SDNMSA and EDNA) rely on neighbouring interactions, but EDNA constructs the substitution matrix using Boltzmann distributions of the dinucleotides based on stacking free energy. Both methods were tested

using the gapped prediction method devised previously versus other conventional alignment tools (ClustalW and Dialign) and the ungapped method. The superiority of both alignment methods over conventional alignment methods was apparent from enhanced prediction power across 18 of the global binding sites in *E. coli*, with EDNA proving to be optimal. Finally, I presented a new alignment colouring, which uses the base stacking free energy as the base for assigning a colour for every nucleotide (Chapter 3).

To further enhance the prediction power, I considered using ChIP-on-chip signal along with likelihood to filter binding sites. First, I established the presence of linear correlation between regions in ChIP-on-chip signals and statistical likelihood. Second, I optimized two intersecting cut-offs for both likelihood values and ChIP-on-chip signals using the known binding sites as true positives. For the two binding sites were considered, LexA and CRP, the prediction of binding sites using both methods have been found to provide better accuracy for conserved binding sites as LexA rather than promiscuous binding sites as CRP (Chapter 4).

In the second part of the thesis, to enhance the modelling of gene regulatory networks, I have introduced a new modelling language that introduces a new modelling paradigm using the chemical compound as an encapsulated modelling unit. In this language, I have also introduced many other features that are missed in other languages, such as reusability where the modeller can import models previously written and reuse chemical compounds already defined. I have also introduced extensions, where the chemical compound reactions can be extended to another chemical compound that redefines some of its reactions. The focus on chemical compounds being the modelling units shifts the modeller from the conventional chemical reaction as

modelling units as in SBML, to encapsulated chemical compounds that can be reused and extended.

Generally, for the past two decades, genome research has transformed our understanding of biological systems. I have argued in this thesis that, with the recent advancement of sequencing technologies, the quantity of molecular data about species in nature is becoming immense, resulting in increased demand for faster methods to analyze such data. I have aimed to provide a better understanding of biological systems, both through the discovering of gene regulatory networks by enhancing binding site prediction, and through the modelling of network dynamics through enhancing modelling languages used. In summary, computational involvement in genome, research is increasingly important, and the work done in this thesis is a manifestation of this fact.

# Chapter 7

# REFERENCES


454-SEQUENCING. 2011. *http://www.454.com* [Online]. [Accessed 21/09 2011].

A. FUNAHASHI , M. MOROHASHI, H. KITANO & TANIMURA, N. 2003. Cell Designer: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO,* Volume 1**,** 159.

A. PHILLIPS & CARDELLI, L. 2004. A Correct Abstract Machine for the Stochastic Pi-calculus. *Workshop on Concurrent Models in Molecular Biology - BioConcur.* London.

AGARWAL P., B. V. Year. Detecting non-adjoining correlations with signals in DNA. *In:* Annual Conference on Research in Computational Molecular Biology, March 22 - 25, 1998 1998 New York, United States. ACM, 2-8.

ALLAWI, H. T. & SANTALUCIA, J., JR. 1997. Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry,* 36**,** 10581-94.

ALLAWI, H. T. & SANTALUCIA, J., JR. 1998a. Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects. *Biochemistry,* 37**,** 9435-44.

ALLAWI, H. T. & SANTALUCIA, J., JR. 1998b. Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. *Biochemistry,* 37**,** 2170-9.

ALLAWI, H. T. & SANTALUCIA, J., JR. 1998c. NMR solution structure of a DNA dodecamer containing single G.T mismatches. *Nucleic Acids Res,* 26**,** 4925-34.

ALLAWI, H. T. & SANTALUCIA, J., JR. 1998d. Thermodynamics of internal C.T mismatches in DNA. *Nucleic Acids Res,* 26**,** 2694-701.

ALON, U. 2006. *An Introduction to Systems Biology - Design Principles of Biological Circuits*, Chapman and Hall/CRC.

ALTSCHUL, S. F. & ERICKSON, B. W. 1986. Optimal sequence alignment using affine gap costs. *Bull Math Biol,* 48**,** 603-16.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

APARICIO, O., GEISBERG, J. V., SEKINGER, E., YANG, A., MOQTADERI, Z. & STRUHL, K. 2005. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr Protoc Mol Biol,* Chapter 21**,** Unit 21 3.

ARAUZO-BRAVO, M. J., FUJII, S., KONO, H., AHMAD, S. & SARAI, A. 2005. Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: toward understanding the indirect readout mechanism in protein-DNA recognition. *J Am Chem Soc,* 127**,** 16074-89.

ARNOSTI, D. N. & KULKARNI, M. M. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem,* 94**,** 890-8.

ATLAS, J. C., NIKOLAEV, E. V., BROWNING, S. T. & SHULER, M. L. 2008. Incorporating genome-wide DNA sequence information into a dynamic whole-cell model of Escherichia coli: application to DNA replication. *IET Syst Biol,* 2**,** 369-82.

BABU, M. M., LUSCOMBE, N. M., ARAVIND, L., GERSTEIN, M. & TEICHMANN, S. A. 2004. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol,* 14**,** 283-91.

BAILEY, T. L. & ELKAN, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol,* 2**,** 28-36.

BARASH Y, E. G., FRIEDMAN N, KAPLAN T. Year. Modeling Dependencies in Protein-DNA Binding Sites. *In:* Proceedings of the 7th International Conference on Research in Computational Molecular Biology (RECOMB), 2003 Berlin. 28-37.

BARENDREGT, H. P. 1984. *The lambda calculus : its syntax and semantics*, North-Holland ; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co.

BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T. Y., SCHONES, D. E., WANG, Z., WEI, G., CHEPELEV, I. & ZHAO, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell,* 129**,** 823-37.

BARTEL, D. P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell,* 116**,** 281-97.

BARTEL, D. P. 2009. MicroRNAs: target recognition and regulatory functions. *Cell,* 136**,** 215-33.

BECKER, S. A., FEIST, A. M., MO, M. L., HANNUM, G., PALSSON, B. O. & HERRGARD, M. J. 2007. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc,* 2**,** 727-38.

BEN-GAL, I., SHANI, A., GOHR, A., GRAU, J., ARVIV, S., SHMILOVICI, A., POSCH, S. & GROSSE, I. 2005. Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics,* 21**,** 2657-66.

BENOUKRAF, T., CAUCHY, P., FENOUIL, R., JEANNIARD, A., KOCH, F., JAEGER, S., THIEFFRY, D., IMBERT, J., ANDRAU, J. C., SPICUGLIA, S. & FERRIER, P. 2009. CoCAS: a ChIP-on-chip analysis suite. *Bioinformatics,* 25**,** 954-5.

BENTLEY, D. R., BALASUBRAMANIAN, S., SWERDLOW, H. P., SMITH, G. P., MILTON, J., BROWN, C. G., HALL, K. P., EVERS, D. J., BARNES, C. L., BIGNELL, H. R., BOUTELL, J. M., BRYANT, J., CARTER, R. J., KEIRA CHEETHAM, R., COX, A. J., ELLIS, D. J., FLATBUSH, M. R., GORMLEY, N. A., HUMPHRAY, S. J., IRVING, L. J., KARBELASHVILI, M. S., KIRK, S. M., LI, H., LIU, X., MAISINGER, K. S., MURRAY, L. J., OBRADOVIC, B., OST, T., PARKINSON, M. L., PRATT, M. R., RASOLONJATOVO, I. M., REED, M. T., RIGATTI, R., RODIGHIERO, C., ROSS, M. T., SABOT, A., SANKAR, S. V., SCALLY, A., SCHROTH, G. P., SMITH, M. E., SMITH, V. P., SPIRIDOU, A., TORRANCE, P. E., TZONEV, S. S., VERMAAS, E. H., WALTER, K., WU, X., ZHANG, L., ALAM, M. D., ANASTASI, C., ANIEBO, I. C., BAILEY, D. M., BANCARZ, I. R., BANERJEE, S., BARBOUR, S. G., BAYBAYAN, P. A., BENOIT, V. A., BENSON, K. F., BEVIS, C., BLACK, P. J., BOODHUN, A., BRENNAN, J. S., BRIDGHAM, J. A., BROWN, R. C., BROWN, A. A., BUERMANN, D. H., BUNDU, A. A., BURROWS, J. C., CARTER, N. P., CASTILLO, N., CHIARA, E. C. M., CHANG, S., NEIL COOLEY, R., CRAKE, N. R., DADA, O. O., DIAKOUMAKOS, K. D., DOMINGUEZ-FERNANDEZ, B., EARNSHAW, D. J., EGBUJOR, U. C., ELMORE, D. W., ETCHIN, S. S., EWAN, M. R., FEDURCO, M., FRASER, L. J., FUENTES FAJARDO, K. V., SCOTT FUREY, W., GEORGE, D., GIETZEN, K. J., GODDARD, C. P., GOLDA, G. S., GRANIERI, P. A., GREEN, D. E., GUSTAFSON, D. L., HANSEN, N. F., HARNISH, K., HAUDENSCHILD, C. D., HEYER, N. I., HIMS, M. M., HO, J. T., HORGAN, A. M., et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456**,** 53-9.

BERG, J. M., TYMOCZKO, J. L. & STRYER, L. 2007. *Biochemistry,* New York, W.H. Freeman.

BIRD, A. 2007. Perceptions of epigenetics. *Nature,* 447**,** 396-8.

BRAY, N., DUBCHAK, I. & PACHTER, L. 2003. AVID: A global alignment program. *Genome Res,* 13**,** 97-102.

BRITTEN, R. J. & DAVIDSON, E. H. 1969. Gene regulation for higher cells: a theory. *Science,* 165**,** 349-57.

BROWN, W. M., PRAGER, E. M., WANG, A. & WILSON, A. C. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J Mol Evol,* 18**,** 225-39.

BRUDNO, M., CHAPMAN, M., GOTTGENS, B., BATZOGLOU, S. & MORGENSTERN, B. 2003a. Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics,* 4**,** 66.

BRUDNO, M., DO, C. B., COOPER, G. M., KIM, M. F., DAVYDOV, E., GREEN, E. D., SIDOW, A. & BATZOGLOU, S. 2003b. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res,* 13**,** 721-31.

BRUDNO, M., MALDE, S., POLIAKOV, A., DO, C. B., COURONNE, O., DUBCHAK, I. & BATZOGLOU, S. 2003c. Glocal alignment: finding rearrangements during alignment. *Bioinformatics,* 19 Suppl 1**,** i54-62.

BULYK, M. L., JOHNSON, P. L. & CHURCH, G. M. 2002. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res,* 30**,** 1255-61.

CALDER, M., GILMORE, S. & HILLSTON, J. 2006. Modelling the influence of RKIP on the ERK signalling pathway using the stochastic process algebra PEPA. *Transactions on Computational Systems Biology,* Volume II**,** 1-23.

CALLADINE, C. R. & DREW, H. R. 1986. Principles of sequence-dependent flexure of DNA. *J Mol Biol,* 192**,** 907-18.

CALZONE, L., FAGES, F. & SOLIMAN, S. 2006. BIOCHAM: an environment for modeling biological systems and formalizing experimental knowledge. *Bioinformatics,* 22**,** 1805-7.

CAO, T. I. D. A. L. A. 2009. Computing Substitution Matrices for Genomic Comparative Analysis. *ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING,* 5476/2009**,** 647-655.

CARROLL, H., BECKSTEAD, W., O'CONNOR, T., EBBERT, M., CLEMENT, M., SNELL, Q. & MCCLELLAN, D. 2007. DNA reference alignment benchmarks based on tertiary structure of encoded proteins. *Bioinformatics,* 23**,** 2648-9.

CHOUDHURI 2004. Gene Regulation and Molecular Toxicology. *Toxicology Mechanisms and Methods,* 15**,** 1-23.

CHOUINARD, J.-Y., FORTIER, P., GULLIVER, T. A. & SPRINGERLINK (ONLINE SERVICE) 1996. *Information theory and applications II 4th Canadian workshop, Lac Delage, Quâebec, Canada, May 28-30, 1995 : selected papers,* Berlin ; New York, Springer.

CROOKS, G. E., HON, G., CHANDONIA, J. M. & BRENNER, S. E. 2004. WebLogo: a sequence logo generator. *Genome Res,* 14**,** 1188-90.

CURIS, E., NICOLIS, I., BENSACI, J., DESCHAMPS, P. & BENAZETH, S. 2009. Mathematical modeling in metal metabolism: overview and perspectives. *Biochimie,* 91**,** 1238-54.

CURTIS, S. E. & CLEGG, M. T. 1984. Molecular evolution of chloroplast DNA sequences. *Mol Biol Evol,* 1**,** 291-301.

DAYHOFF, M. O. 1978. A model of Evolutionary Change in Proteins. *Atlas of protein sequence and structure,* 5**,** 345–358.

DELCHER, A. L., KASIF, S., FLEISCHMANN, R. D., PETERSON, J., WHITE, O. & SALZBERG, S. L. 1999. Alignment of whole genomes. *Nucleic Acids Res,* 27**,** 2369-76.

DJORDJEVIC, M., SENGUPTA, A. M. & SHRAIMAN, B. I. 2003. A biophysical approach to transcription factor binding site discovery. *Genome Res,* 13**,** 2381-90.

DURBIN, EDDY, S. R., KROGH, A. & MITCHISON, G. 1998. *Biological sequence analysis : probabilistic models of proteins and nucleic acids,* Cambridge, UK New York, Cambridge University Press.

EDDY, S. R. 2004. Where did the BLOSUM62 alignment score matrix come from? *Natue Biotechnology,* 22**,** 1.

EDGAR, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* 5**,** 113.

ERRAMPALLI, D. D., PRIAMI, C. & QUAGLIA, P. 2004. A formal language for computational systems biology. *OMICS,* 8**,** 370-80.

FAUTEUX, F., BLANCHETTE, M. & STROMVIK, M. V. 2008. Seeder: discriminative seeding DNA motif discovery. *Bioinformatics,* 24**,** 2303-7.

FINNEY, A. & HUCKA, M. 2003. Systems biology markup language: Level 2 and beyond. *Biochem Soc Trans,* 31**,** 1472-3.

FONG, S. S., BURGARD, A. P., HERRING, C. D., KNIGHT, E. M., BLATTNER, F. R., MARANAS, C. D. & PALSSON, B. O. 2005. In silico design and adaptive evolution of Escherichia coli for production of lactic acid. *Biotechnol Bioeng,* 91**,** 643-8.

FONG, S. S., MARCINIAK, J. Y. & PALSSON, B. O. 2003. Description and interpretation of adaptive evolution of Escherichia coli K-12 MG1655 by using a genome-scale in silico metabolic model. *J Bacteriol,* 185**,** 6400-8.

FRIBERG, M., VON ROHR, P. & GONNET, G. 2005. Scoring functions for transcription factor binding site prediction. *BMC Bioinformatics,* 6**,** 84.

GALLEGOS, M. T., SCHLEIF, R., BAIROCH, A., HOFMANN, K. & RAMOS, J. L. 1997. Arac/XylS family of transcriptional regulators. *Microbiol Mol Biol Rev,* 61**,** 393-410.

GAMA-CASTRO, S., JIMENEZ-JACINTO, V., PERALTA-GIL, M., SANTOS-ZAVALETA, A., PENALOZA-SPINOLA, M. I., CONTRERAS-MOREIRA, B., SEGURA-SALAZAR, J., MUNIZ-RASCADO, L., MARTINEZ-FLORES, I., SALGADO, H., BONAVIDES-MARTINEZ, C., ABREU-GOODGER, C., RODRIGUEZ-PENAGOS, C., MIRANDA-RIOS, J., MORETT, E., MERINO, E., HUERTA, A. M., TREVINO-QUINTANILLA, L. & COLLADO-VIDES, J. 2008. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res,* 36**,** D120-4.

GENTLEMAN, R. C., CAREY, V. J., BATES, D. M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., HORNIK, K., HOTHORN, T., HUBER, W., IACUS, S., IRIZARRY, R., LEISCH, F., LI, C., MAECHLER, M., ROSSINI, A. J., SAWITZKI, G., SMITH, C., SMYTH, G., TIERNEY, L., YANG, J. Y. & ZHANG, J. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol,* 5**,** R80.

GHEORGHE, M. & MITRANA, V. 2004. A formal language-based approach in biology. *Comp Funct Genomics,* 5**,** 91-4.

GILLESPIE 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys,* 22**,** 403–34.

GOJOBORI, T., LI, W. H. & GRAUR, D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. *J Mol Evol,* 18**,** 360-9.

GOSS, P. J. & PECCOUD, J. 1998. Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. *Proc Natl Acad Sci U S A,* 95**,** 6750-5.

GRAINGER, D. C., HURD, D., HARRISON, M., HOLDSTOCK, J. & BUSBY, S. J. 2005. Studies of the distribution of Escherichia coli cAMP-receptor protein and RNA polymerase along the E. coli chromosome. *Proc Natl Acad Sci U S A,* 102**,** 17693-8.

GROMIHA, M. M., SIEBERS, J. G., SELVARAJ, S., KONO, H. & SARAI, A. 2005. Role of inter and intramolecular interactions in protein-DNA recognition. *Gene,* 364**,** 108-13.

GRUBER, T. M. & GROSS, C. A. 2003. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol,* 57**,** 441-66.

GUERRIERO, JOHN K. HEATH & PRIAMI, C. 2007. An Automated Translation from a Narrative Language for Biological Modelling into Process Algebra. *COMPUTATIONAL METHODS IN SYSTEMS BIOLOGY.*

HALL, N. 2007. Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol,* 210**,** 1518-25.

HE, L. & HANNON, G. J. 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet,* 5**,** 522-31.

HELMANN, J. D. & CHAMBERLIN, M. J. 1988. Structure and function of bacterial sigma factors. *Annu Rev Biochem,* 57**,** 839-72.

HENIKOFF, S. & HENIKOFF, J. G. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Res,* 19**,** 6565-72.

HENIKOFF, S. & HENIKOFF, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A,* 89**,** 10915-9.

HERTZ, G. Z. & STORMO, G. D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics,* 15**,** 563-77.

HOCHBERG, Y. B. Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological),* 57**,** 289-300.

HOLLIDAY, R. 1990. Mechanisms for the control of gene activity during development. *Biol Rev Camb Philos Soc,* 65**,** 431-71.

HOOPS, S., SAHLE, S., GAUGES, R., LEE, C., PAHLE, J., SIMUS, N., SINGHAL, M., XU, L., MENDES, P. & KUMMER, U. 2006. COPASI--a COmplex PAthway SImulator. *Bioinformatics,* 22**,** 3067-74.

HUA, Q., JOYCE, A. R., FONG, S. S. & PALSSON, B. O. 2006. Metabolic analysis of adaptive evolution for in silico-designed lactate-producing strains. *Biotechnol Bioeng,* 95**,** 992-1002.

HUBLITZ, P., ALBERT, M. & PETERS, A. H. 2009. Mechanisms of transcriptional repression by histone lysine methylation. *Int J Dev Biol,* 53**,** 335-54.

HUGHES, J. D., ESTEP, P. W., TAVAZOIE, S. & CHURCH, G. M. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol,* 296**,** 1205-14.

IBARRA, R. U., EDWARDS, J. S. & PALSSON, B. O. 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature,* 420**,** 186-9.

INA, Y. 1998. Estimation of the transition/transversion ratio. *J Mol Evol,* 46**,** 521-33.

JACOB, F. & MONOD, J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol,* 3**,** 318-56.

JENUWEIN, T. & ALLIS, C. D. 2001. Translating the histone code. *Science,* 293**,** 1074-80.

JOHNSON, D. S., MORTAZAVI, A., MYERS, R. M. & WOLD, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science,* 316**,** 1497-502.

JUST, W. 2001. Computational complexity of multiple sequence alignment with SP-score. *J Comput Biol,* 8**,** 615-23.

KAHRAMANOĞULLARI, O. C., LUCA; CARON, EMMANUELLE 2009. An Intuitive Automated Modelling Interface for Systems Biology. *Electronic Proceedings in Theoretical Computer Science,* s.

KAPANIDIS, A. N., MARGEAT, E., LAURENCE, T. A., DOOSE, S., HO, S. O., MUKHOPADHYAY, J., KORTKHONJIA, E., MEKLER, V., EBRIGHT, R. H. & WEISS, S. 2005. Retention of transcription initiation factor sigma70 in transcription elongation: single-molecule analysis. *Mol Cell,* 20**,** 347-56.

KATOH, K., KUMA, K., TOH, H. & MIYATA, T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res,* 33**,** 511-8.

KHAN, S. H. & KUMAR, R. 2009. An overview of the importance of conformational flexibility in gene regulation by the transcription factors. *J Biophys,* 2009**,** 210485.

KHOLODENKO, B. N. 2006. Cell-signalling dynamics in time and space. *Nat Rev Mol Cell Biol,* 7**,** 165-76.

KIM, N. K., THARAKARAMAN, K., MARINO-RAMIREZ, L. & SPOUGE, J. L. 2008. Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. *BMC Bioinformatics,* 9**,** 262.

KING, O. D. & ROTH, F. P. 2003. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res,* 31**,** e116.

KOCH, C. M., ANDREWS, R. M., FLICEK, P., DILLON, S. C., KARAOZ, U., CLELLAND, G. K., WILCOX, S., BEARE, D. M., FOWLER, J. C., COUTTET, P., JAMES, K. D., LEFEBVRE, G. C., BRUCE, A. W., DOVEY, O. M., ELLIS, P. D., DHAMI, P., LANGFORD, C. F., WENG, Z., BIRNEY, E., CARTER, N. P., VETRIE, D. & DUNHAM, I. 2007. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res,* 17**,** 691-707.

LAGOS-QUINTANA, M., RAUHUT, R., LENDECKEL, W. & TUSCHL, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science,* 294**,** 853-8.

LANAVE, C., TOMMASI, S., PREPARATA, G. & SACCONE, C. 1986. Transition and transversion rate in the evolution of animal mitochondrial DNA. *Biosystems,* 19**,** 273-83.

LATCHMAN, D. S. 1997. Transcription factors: an overview. *Int J Biochem Cell Biol,* 29**,** 1305-12.

LAURIA, M., IORIO, F. & DI BERNARDO, D. 2009. NIRest: a tool for gene network and mode of action inference. *Ann N Y Acad Sci,* 1158**,** 257-64.

LAWRENCE, C. E., ALTSCHUL, S. F., BOGUSKI, M. S., LIU, J. S., NEUWALD, A. F. & WOOTTON, J. C. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science,* 262**,** 208-14.

LAWRENCE, C. E. & REILLY, A. A. 1990. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins,* 7**,** 41-51.

LEBLANC, B. & MOSS, T. 2001. DNase I footprinting. *Methods Mol Biol,* 148**,** 31-8.

LEVENSHTEIN 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady,* 10**,** 707–10.

LI, J., WANG, L., HASHIMOTO, Y., TSAO, C. Y., WOOD, T. K., VALDES, J. J., ZAFIRIOU, E. & BENTLEY, W. E. 2006. A stochastic model of Escherichia coli AI-2 quorum signal circuit reveals alternative synthesis pathways. *Mol Syst Biol,* 2**,** 67.

LI, X. Y., MACARTHUR, S., BOURGON, R., NIX, D., POLLARD, D. A., IYER, V. N., HECHMER, A., SIMIRENKO, L., STAPLETON, M., LUENGO HENDRIKS, C. L., CHU, H. C., OGAWA, N., INWOOD, W., SEMENTCHENKO, V., BEATON, A., WEISZMANN, R., CELNIKER, S. E., KNOWLES, D. W., GINGERAS, T., SPEED, T. P., EISEN, M. B. & BIGGIN, M. D. 2008. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol,* 6**,** e27.

LIU, J. S. & LAWRENCE, C. E. 1999. Bayesian inference on biopolymer models. *Bioinformatics,* 15**,** 38-52.

LOPES, S., NEVES, C. S., EATON, P. & GAMEIRO, P. Cardiolipin, a key component to mimic the E. coli bacterial membrane in model systems revealed by dynamic light scattering and steady-state fluorescence anisotropy. *Anal Bioanal Chem,* 398**,** 1357-66.

MACHNE, R., FINNEY, A., MULLER, S., LU, J., WIDDER, S. & FLAMM, C. 2006. The SBML ODE Solver Library: a native API for symbolic and fast numerical analysis of reaction networks. *Bioinformatics,* 22**,** 1406-7.

MARDIS, E. R. 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet,* 9**,** 387-402.

MARTIN, D. I. & ORKIN, S. H. 1990. Transcriptional activation and DNA binding by the erythroid factor GF-1/NF-E1/Eryf 1. *Genes Dev,* 4**,** 1886-98.

MARTINEZ-ANTONIO, A. & COLLADO-VIDES, J. 2003. Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol,* 6**,** 482-9.

MARTINEZ, M. B., FLICKINGER, M. C. & NELSESTUEN, G. L. 1999. Steady-state enzyme kinetics in the Escherichia coli periplasm: a model of a whole cell biocatalyst. *J Biotechnol,* 71**,** 59-66.

MATHEWS, D. H., SABINA, J., ZUKER, M. & TURNER, D. H. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol,* 288**,** 911-40.

MATSUNO, Y. H., A. DOI, S. MIYANO 2002. Genomic Object Net: On-going report on biopathway modeling and simulation. *Currents in Computational Molecular Biology***,** 132-133.

MC GRATH, S. & SINDEREN, D. V. 2007. *Bacteriophage : genetics and molecular biology,* Norfolk, UK, Caister Academic Press.

MCKERNAN, K. J., PECKHAM, H. E., COSTA, G. L., MCLAUGHLIN, S. F., FU, Y., TSUNG, E. F., CLOUSER, C. R., DUNCAN, C., ICHIKAWA, J. K., LEE, C. C., ZHANG, Z., RANADE, S. S., DIMALANTA, E. T., HYLAND, F. C., SOKOLSKY, T. D., ZHANG, L., SHERIDAN, A., FU, H., HENDRICKSON, C. L., LI, B., KOTLER, L., STUART, J. R., MALEK, J. A., MANNING, J. M., ANTIPOVA, A. A., PEREZ, D. S., MOORE, M. P., HAYASHIBARA, K. C., LYONS, M. R., BEAUDOIN, R. E., COLEMAN, B. E., LAPTEWICZ, M. W., SANNICANDRO, A. E., RHODES, M. D., GOTTIMUKKALA, R. K., YANG, S., BAFNA, V., BASHIR, A., MACBRIDE, A., ALKAN, C., KIDD, J. M., EICHLER, E. E., REESE, M. G., DE LA VEGA, F. M. & BLANCHARD, A. P. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res,* 19**,** 1527-41.

MERKULOVA, T. I., OSHCHEPKOV, D. Y., IGNATIEVA, E. V., ANANKO, E. A., LEVITSKY, V. G., VASILIEV, G. V., KLIMOVA, N. V., MERKULOV, V. M. & KOLCHANOV, N. A. 2007. Bioinformatical and experimental approaches to investigation of transcription factor binding sites in vertebrate genes. *Biochemistry (Mosc),* 72**,** 1187-93.

MORGENSTERN, B. 2007. Alignment of genomic sequences using DIALIGN. *Methods Mol Biol,* 395**,** 195-204.

MOSES, A. M., CHIANG, D. Y., KELLIS, M., LANDER, E. S. & EISEN, M. B. 2003. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol,* 3**,** 19.

MOSES, A. M., POLLARD, D. A., NIX, D. A., IYER, V. N., LI, X. Y., BIGGIN, M. D. & EISEN, M. B. 2006. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol,* 2**,** e130.

NEEDLEMAN, S. B. & 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**,** 443-53.

NIKEL, P. I., PETTINARI, M. J., RAMIREZ, M. C., GALVAGNO, M. A. & MENDEZ, B. S. 2008. Escherichia coli arcA mutants: metabolic profile characterization of microaerobic cultures using glycerol as a carbon source. *J Mol Microbiol Biotechnol,* 15**,** 48-54.

NOTREDAME, C., HIGGINS, D. G. & HERINGA, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol,* 302**,** 205-17.

OSADA, R., ZASLAVSKY, E. & SINGH, M. 2004. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics,* 20**,** 3516-25.

PAVESI, G., MAURI, G. & PESOLE, G. 2001. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics,* 17 Suppl 1**,** S207-14.

PFEIFFER, T., SANCHEZ-VALDENEBRO, I., NUNO, J. C., MONTERO, F. & SCHUSTER, S. 1999. METATOOL: for studying metabolic networks. *Bioinformatics,* 15**,** 251-7.

PONTING, C. P., SCHULTZ, J., MILPETZ, F. & BORK, P. 1999. SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res,* 27**,** 229-32.

PRESSER, K. A., ROSS, T. & RATKOWSKY, D. A. 1998. Modelling the growth limits (growth/no growth interface) of Escherichia coli as a function of temperature, pH, lactic acid concentration, and water activity. *Appl Environ Microbiol,* 64**,** 1773-9.

PRIAMI , REGEV, A., W. SILVERMAN & SHAPIRO, E. 2001. Application of a stochastic name-passing calculus to representation and simulation of molecular processes. *Information Processing Letters,* Volume 80**,** 25, 80.

PRIAMI, C. 2005. Beta Binders for Biological Interactions. *COMPUTATIONAL METHODS IN SYSTEMS BIOLOGY,* 3082**,** 20-33.

PURVIS, A. & BROMHAM, L. 1997. Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny. *J Mol Evol,* 44**,** 112-9.

RAGHAVA, G. P., SEARLE, S. M., AUDLEY, P. C., BARBER, J. D. & BARTON, G. J. 2003. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics,* 4**,** 47.

REDHEAD, E. & BAILEY, T. L. 2007. Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics,* 8**,** 385.

REN, B., ROBERT, F., WYRICK, J. J., APARICIO, O., JENNINGS, E. G., SIMON, I., ZEITLINGER, J., SCHREIBER, J., HANNETT, N., KANIN, E., VOLKERT, T. L., WILSON, C. J., BELL, S. P. & YOUNG, R. A. 2000. Genome-wide location and function of DNA binding proteins. *Science,* 290**,** 2306-9.

RONAGHI, M., UHLEN, M. & NYREN, P. 1998. A sequencing method based on real-time pyrophosphate. *Science,* 281**,** 363, 365.

ROSENFELD, J. A., WANG, Z., SCHONES, D. E., ZHAO, K., DESALLE, R. & ZHANG, M. Q. 2009. Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics,* 10**,** 143.

ROSS, T., RATKOWSKY, D. A., MELLEFONT, L. A. & MCMEEKIN, T. A. 2003. Modelling the effects of temperature, water activity, pH and lactic acid concentration on the growth rate of Escherichia coli. *Int J Food Microbiol,* 82**,** 33-43.

RUSSO, V. E. A., MARTIENSSEN, R. A. & RIGGS, A. D. 1996. *Epigenetic mechanisms of gene regulation,* Plainview, N.Y., Cold Spring Harbor Laboratory Press.

SAITOU, N. & NEI, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol,* 4**,** 406-25.

SALAMA, R. A. & STEKEL, D. J. 2010. Inclusion of neighboring base interdependencies substantially improves genome-wide prokaryotic transcription factor binding site prediction. *Nucleic Acids Res,* 38**,** e135.

SANGIORGI, D. & WALKER, D. 2001. *The [pi]-calculus : a theory of mobile processes,* Cambridge, [England] ; New York, Cambridge University Press.

SANTALUCIA, J., JR. & TURNER, D. H. 1997. Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers,* 44**,** 309-19.

SARAI, A. & KONO, H. 2005. Protein-DNA recognition patterns and predictions. *Annu Rev Biophys Biomol Struct,* 34**,** 379-98.

SARAI, A., SIEBERS, J., SELVARAJ, S., GROMIHA, M. M. & KONO, H. 2005. Integration of bioinformatics and computational biology to understand protein-DNA recognition mechanism. *J Bioinform Comput Biol,* 3**,** 169-83.

SAURO, F. B. A. H. 2006. Human-readable model defnition language. *Technical report, Keck Graduate Institute*.

SAURO, H. M., HUCKA, M., FINNEY, A., WELLOCK, C., BOLOURI, H., DOYLE, J. & KITANO, H. 2003. Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS,* 7**,** 355-72.

SCHLITT, T. & BRAZMA, A. 2005. Modelling gene networks at different organisational levels. *FEBS Lett,* 579**,** 1859-66.

SCHWARTZ, S., ZHANG, Z., FRAZER, K. A., SMIT, A., RIEMER, C., BOUCK, J., GIBBS, R., HARDISON, R. & MILLER, W. 2000. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res,* 10**,** 577-86.

SEDWARDS, S. & MAZZA, T. 2007. Cyto-Sim: a formal language model and stochastic simulator of membrane-enclosed biochemical processes. *Bioinformatics,* 23**,** 2800-2.

SESHASAYEE, A. S., SIVARAMAN, K. & LUSCOMBE, N. M. 2011. An overview of prokaryotic transcription factors : a summary of function and occurrence in bacterial genomes. *Subcell Biochem,* 52**,** 7-23.

SHARMA, U. K. & CHATTERJI, D. 2010. Transcriptional switching in Escherichia coli during stress and starvation by modulation of sigma activity. *FEMS Microbiol Rev,* 34**,** 646-57.

SIDDHARTHAN, R. 2006. Sigma: multiple alignment of weakly-conserved non-coding DNA sequence. *BMC Bioinformatics,* 7**,** 143.

SIDDHARTHAN, R. 2008. PhyloGibbs-MP: module prediction and discriminative motif-finding by Gibbs sampling. *PLoS Comput Biol,* 4**,** e1000156.

SIDDHARTHAN, R., SIGGIA, E. D. & VAN NIMWEGEN, E. 2005. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol,* 1**,** e67.

SIDDHARTHAN, R. & VAN NIMWEGEN, E. 2007. Detecting regulatory sites using PhyloGibbs. *Methods Mol Biol,* 395**,** 381-402.

SMITH, A. D., SUMAZIN, P. & ZHANG, M. Q. 2005. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A,* 102**,** 1560-5.

SMITH, R. F. S. A. T. F. 1992. Pattern-induced multi-sequence alignment (PUMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Engineering* 5**,** 6.

SMITH, T. F. & WATERMAN, M. S. 1981. Identification of common molecular subsequences. *J Mol Biol,* 147**,** 195-7.

SMOLEN, P., BAXTER, D. A. & BYRNE, J. H. 2000. Modeling transcriptional control in gene networks--methods, recent results, and future directions. *Bull Math Biol,* 62**,** 247-92.

SOKAL R, M. C. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38**,** 29.

SPILIANAKIS, C. G., LALIOTI, M. D., TOWN, T., LEE, G. R. & FLAVELL, R. A. 2005. Interchromosomal associations between alternatively expressed loci. *Nature,* 435**,** 637-45.

SRIVASTAVA, R., PETERSON, M. S. & BENTLEY, W. E. 2001. Stochastic kinetic analysis of the Escherichia coli stress circuit using sigma(32)-targeted antisense. *Biotechnol Bioeng,* 75**,** 120-9.

STORMO, G. D., SCHNEIDER, T. D. & GOLD, L. 1986. Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res,* 14**,** 6661-79.

STORMS, V., CLAEYS, M., SANCHEZ, A., DE MOOR, B., VERSTUYF, A. & MARCHAL, K. 2010. The effect of orthology and coregulation on detecting regulatory motifs. *PLoS One,* 5**,** e8938.

STRAHL, B. D. & ALLIS, C. D. 2000. The language of covalent histone modifications. *Nature,* 403**,** 41-5.

STRANDBERG, A. K. & SALTER, L. A. 2004. A comparison of methods for estimating the transition:transversion ratio from DNA sequences. *Mol Phylogenet Evol,* 32**,** 495-503.

STROHMAN, R. 2002. Maneuvering in the complex path from genotype to phenotype. *Science,* 296**,** 701-3.

SWINNEN, I. A., BERNAERTS, K. & VAN IMPE, J. F. 2006. Modelling the work to be done by Escherichia coli to adapt to sudden temperature upshifts. *Lett Appl Microbiol,* 42**,** 507-13.

TEODORESCU, O., GALOR, T., PILLARDY, J. & ELBER, R. 2004. Enriching the sequence substitution matrix by structural information. *Proteins,* 54**,** 41-8.

THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics,* Chapter 2**,** Unit 2 3.

THOMPSON, J. D., KOEHL, P., RIPP, R. & POCH, O. 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins,* 61**,** 127-36.

THOMPSON, J. D., PLEWNIAK, F. & POCH, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res,* 27**,** 2682-90.

THOMPSON, W., ROUCHKA, E. C. & LAWRENCE, C. E. 2003. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res,* 31**,** 3580-5.

THOMPSON, W. A., NEWBERG, L. A., CONLAN, S., MCCUE, L. A. & LAWRENCE, C. E. 2007. The Gibbs Centroid Sampler. *Nucleic Acids Res,* 35**,** W232-7.

TOEDLING, J., SKYLAR, O., KRUEGER, T., FISCHER, J. J., SPERLING, S. & HUBER, W. 2007. Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics,* 8**,** 221.

TOMITA, M., HASHIMOTO, K., TAKAHASHI, K., SHIMIZU, T. S., MATSUZAKI, Y., MIYOSHI, F., SAITO, K., TANIDA, S., YUGI, K., VENTER, J. C. & HUTCHISON, C. A., 3RD 1999. E-CELL: software environment for whole-cell simulation. *Bioinformatics,* 15**,** 72-84.

TOMOVIC, A. & OAKELEY, E. J. 2007. Position dependencies in transcription factor binding sites. *Bioinformatics,* 23**,** 933-41.

TOMPA, M., LI, N., BAILEY, T. L., CHURCH, G. M., DE MOOR, B., ESKIN, E., FAVOROV, A. V., FRITH, M. C., FU, Y., KENT, W. J., MAKEEV, V. J., MIRONOV, A. A., NOBLE, W. S., PAVESI, G., PESOLE, G., REGNIER, M., SIMONIS, N., SINHA, S., THIJS, G., VAN HELDEN, J., VANDENBOGAERT, M., WENG, Z., WORKMAN, C., YE, C. & ZHU, Z. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol,* 23**,** 137-44.

TYAGI, M., GOWRI, V. S., SRINIVASAN, N., DE BREVERN, A. G. & OFFMANN, B. 2006. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. *Proteins,* 65**,** 32-9.

VASS, M., ALLEN, N., SHAFFER, C. A., RAMAKRISHNAN, N., WATSON, L. T. & TYSON, J. J. 2004. the JigCell model builder and run manager. *Bioinformatics,* 20**,** 3680-1.

VILLEGER, A. C., PETTIFER, S. R. & KELL, D. B. 2010. Arcadia: a visualization tool for metabolic pathways. *Bioinformatics,* 26**,** 1470-1.

VINGRON, M. & WATERMAN, M. S. 1994. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J Mol Biol,* 235**,** 1-12.

WADE, J. T., REPPAS, N. B., CHURCH, G. M. & STRUHL, K. 2005. Genomic analysis of LexA binding reveals the permissive nature of the Escherichia coli genome and identifies unconventional target sites. *Genes Dev,* 19**,** 2619-30.

WAHDE, M. & HERTZ, J. 2001. Modeling genetic regulatory dynamics in neural development. *J Comput Biol,* 8**,** 429-42.

WAKELEY, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *J Mol Evol,* 37**,** 613-23.

WAKELEY, J. 1994. Substitution-rate variation among sites and the estimation of transition bias. *Mol Biol Evol,* 11**,** 436-42.

WALKER, G. C. 2000. Understanding the complexity of an organism's responses to DNA damage. *Cold Spring Harb Symp Quant Biol,* 65**,** 1-10.

WALLE, I., LASTERS, I. & WYNS, L. 2004. Align-m--a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics,* 20**,** 1428-35.

WANG, L. & JIANG, T. 1994. On the complexity of multiple sequence alignment. *J Comput Biol,* 1**,** 337-48.

WANG, M., MOK, M. W., HARPER, H., LEE, W. H., MIN, J., KNAPP, S., OPPERMANN, U., MARSDEN, B. & SCHAPIRA, M. 2010. Structural genomics of histone tail recognition. *Bioinformatics,* 26**,** 2629-30.

WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M., CLAMP, M. & BARTON, G. J. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-91.

WEBB, K. & WHITE, T. 2006. Cell modeling with reusable agent-based formalisms. *Applied Intelligence,* 24**,** 169-181.

WEBSTER, C., GASTON, K. & BUSBY, S. 1988. Transcription from the Escherichia coli melR promoter is dependent on the cyclic AMP receptor protein. *Gene,* 68**,** 297-305.

WHEELER, T. J. & KECECIOGLU, J. D. 2007. Multiple alignment by aligning alignments. *Bioinformatics,* 23**,** i559-68.

WILLEMOES, M., HOVE-JENSEN, B. & LARSEN, S. 2000. Steady state kinetic model for the binding of substrates and allosteric effectors to Escherichia coli phosphoribosyl-diphosphate synthase. *J Biol Chem,* 275**,** 35408-12.

WU, C. F. J. 1983. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics,* 11**,** 95-103.

YANG, Z. & YODER, A. D. 1999. Estimation of the transition/transversion rate bias and species sampling. *J Mol Evol,* 48**,** 274-83.

ZHANG, M. Q. & MARR, T. G. 1993. A weight array method for splicing signal analysis. *Comput Appl Biosci,* 9**,** 499-509.

ZHANG, W. & SUN, Z. 2008. Random local neighbor joining: a new method for reconstructing phylogenetic trees. *Mol Phylogenet Evol,* 47**,** 117-28.

ZHENG, M., BARRERA, L. O., REN, B. & WU, Y. N. 2007. ChIP-chip: data, model, and analysis. *Biometrics,* 63**,** 787-96.

ZHOU, Q. & LIU, J. S. 2004. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics,* 20**,** 909-16.

ZWEIG, M. H. & CAMPBELL, G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem,* 39**,** 561-77.

# APPENDIX I

In this appendix, I have included the preliminary work done on analysis of binding site specificity through a sequence based approach.

The specificity of transcription factors binding to their cognate target binding site has been extensively studied using protein structure based models (Arauzo-Bravo et al., 2005, Gromiha et al., 2005, Sarai and Kono, 2005, Sarai et al., 2005). It has been shown that specificity results from two major interactions between the binding site and the transcription factor: "direct" read outs, representing the direct interactions between the amino acids and the DNA bases through specific hydrogen bonds; and "indirect" readouts resulting from indirect interactions with the side chains of the protein; the latter process is enabled by conformational changes in the DNA structure (Ben-Gal et al., 2005).

The latest research in the subject relies on a heuristic approach to thread the transcription factors over DNA binding sites (Sarai et al., 2005). This approach uses the statistical potential of the amino acids around DNA bases derived from the known DNA protein complexes in the PDB. While this knowledge base approach has proved to be efficient in providing predictive model for the binding site specificity, it is far from being a simple model to be applied on DNA sequences without considering any protein and DNA structural information, which is due to the nature of the problem being structural based.

On the other hand, the potential of using the DNA base stacking interactions captured by the neighbored correlations have proved to be of important significance to this subject. (Tomovic

and Oakeley, 2007) have proved that high dependencies between the DNA binding sites positions are highly correlated with indirect readouts (i.e. indirect contact between transcription factors amino acids and DNA bases through structural docking for instance or any interaction that doesn't involve actual direct bonding between DNA bases and amino acids). It has also been shown that TFBS interactions with large indirect readouts show high dependencies between neighbouring positions of bases in the DNA binding site, corresponding to topological constraints (Ben-Gal et al., 2005). Accordingly; the interdependencies between the binding site positions provide a strong correlation to DNA conformations which in turn affects the indirect readouts. Hence considering the specificity of those interdependencies might lead to a simpler model to analyze the specificity of binding sites.

## Binding Site Specificity

The hypothesis in this work is that transcription factors in a family sharing the same DNA binding domain structure should incur major similarity in DNA conformations, with associated "indirect" read-outs, while each specific transcription factor in that family will exhibit specific conformations with associated "direct" read-outs. A consequence of this hypothesis is that it should be possible to quantify and utilize protein-specific and protein-family-specific conservations in a multiple sequence alignment of TFBS sequences associated with a transcription factor and closely related protein family members. As a preliminary test of this hypothesis, I have grouped the binding site sequences for which the transcription factors have the same DBD and similar binding site length, in order to indentify positions with protein-specific and protein-family-specific conservations of neighbouring position interactions.

Positions with common distributions are the ones corresponding to DNA flexure across the protein family; other positions can be specific to the binding site. As an example, I have shown neighbouring base interactions across three transcription factor binding sites (LexA, OmpR, PhoB) for the Helix Turn Helix DBD (Figure 1).

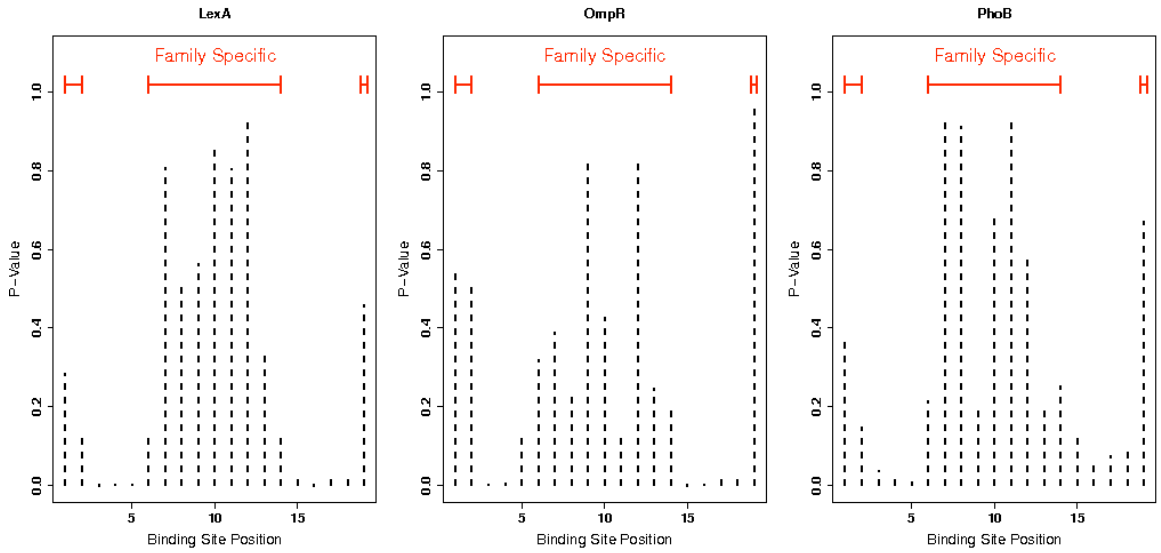


**Figure 1:** represents the specificity of the three binding sites (LexA, OmpR, PhoB) among the HTH family, it shows the significance (y-axis) of each of the neighbouring positions (x-axis) of the binding sites, the higher the p-value on the y-axis the more similar is this position to the family consensus. This figure shows similar signatures between LexA, OmpR and PhoB.

This essentially introduces a new and innovative dimension to the analysis of binding site sequences. Use of this information can enable development of TFBS prediction algorithms for non-global regulators, as a model can be built that includes binding sites of appropriate homologous TFs, without sacrificing specificity. This information will also enable enhanced specificity models even for global regulators.

# APPENDIX II

An abbreviated version of the SBML model presented in section 5.7

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>

<sbml xmlns:xsi="http://www.sbml.org/sbml/level2/version4" level="2"

        xsi:schemaLocation="http://sbml.org/Special/xml-schemas/sbml-l2v4-
schema/sbml.xsd"

        metaid="metaid_0000001" version="4">

        <model id="model01" metaid="metaid_0000002" name="model">

                <listOfCompartments>

                        <compartment id="c_1" metaid="metaid_0000003"
name="Compartment_1" size="1e-14"/>

                </listOfCompartments>

                <listOfUnitDefinitionss>

                        <unitDefinition id="per_second">

                            <listOfUnits>

                               <unit kind="second" exponent="-1"/>

                            </listOfUnits>

                        </unitDefinition>

                        <unitDefinition id="litre_per_mole_per_second">

                            <listOfUnits>

                               <unit kind="mole"   exponent="-1"/>

                               <unit kind="litre"  exponent="1"/>

                               <unit kind="second" exponent="-1"/>
```

```
                </listOfUnits>
            </unitDefinition>
        </listOfUnitDefinitions>
        <listOfSpecies>
            <species compartment="c_1" id="CRP" metaid="metaid_00000030"
                name="CRP" />
            <species compartment="c_1" id="CRPDimer"
metaid="metaid_0000004"
                name="CRPDimer" />
            <species compartment="c_1" id="MelR" metaid="metaid_00000015"
                name="MelR" />
            <species compartment="c_1" id="Melr_Melibiose"
metaid="metaid_00000016"
                name="Melr_Melibiose" />
            <species compartment="c_1" id="melr_CRPCAMP_1"
metaid="metaid_0000009"
                name="melr_CRPCAMP_1" />
            <species compartment="c_1" id="melr_Melrbs2"
metaid="metaid_0000010"
                name="melr_Melrbs2" />
            <species compartment="c_1" id="melr_Melrbs21"
metaid="metaid_0000011"
                name="melr_Melrbs21" />
            <species compartment="c_1" id="melr_Melrbs1"
metaid="metaid_0000012"
                name="melr_Melrbs1" />
```

<species compartment="c_1" id="melr_Melrbs11" metaid="metaid_0000013"

name="melr_Melrbs11" />

<species compartment="c_1" id="melr_CRPCAMP_2" metaid="metaid_0000014"

name="melr_CRPCAMP_2" />

<species compartment="c_1" id="melr_MelrbsR" metaid="metaid_0000017"

name="melr_MelrbsR" />

<species compartment="c_1" id="melr_Melrbs1_MelR" metaid="metaid_0000015"

name="melr_Melrbs1_MelR" />

<species compartment="c_1" id="melr_Melrbs2_MelR" metaid="metaid_0000020"

name="melr_Melrbs2_MelR" />

<species compartment="c_1" id="melr_Melrbs11_MelR" metaid="metaid_0000021"

name="melr_Melrbs11_MelR" />

<species compartment="c_1" id="melr_MelrbsR_MelR" metaid="metaid_0000022"

name="melr_MelrbsR_MelR" />

<species compartment="c_1" id="melr_CRPCAMP_1_CRPDimer"

metaid="metaid_0000023" name="melr_CRPCAMP_1_CRPDimer" />

<species compartment="c_1" id="melr_CRPCAMP_2_CRPDimer"

metaid="metaid_0000024" name="melr_CRPCAMP_2_CRPDimer" />

<species compartment="c_1" id="melr_Melrbs1_Melr_Melibiose" metaid="metaid_0000025"

name="melr_Melrbs1_Melr_Melibiose" />

<species compartment="c_1" id="melr_Melrbs2_Melr_Melibiose" metaid="metaid_0000026"

name="melr_Melrbs2_Melr_Melibiose" />

<species compartment="c_1" id="melr_Melrbs11_Melr_Melibiose" metaid="metaid_0000027"

name="melr_Melrbs11_Melr_Melibiose" />

<species compartment="c_1" id="melr_Melrbs21_Melr_Melibiose" metaid="metaid_0000033"

name="melr_Melrbs21_Melr_Melibiose" />

<species compartment="c_1" id="melr_MelrbsR_Melr_Melibiose" metaid="metaid_0000028"

name="melr_MelrbsR_Melr_Melibiose" />

<species compartment="c_1" id="melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer" metaid="metaid_0000031"

name="melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer" />

<species compartment="c_1" id="melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer_Melrbs2_Melr_Melibiose" metaid="metaid_0000032"

name="melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer_Melrbs2_Melr_Melibiose" />

<species compartment="c_1" id="Melibiose" metaid="metaid_0000029"

name="Melibiose" />

<species compartment="c_1" id="RNA_Polymerase" metaid="metaid_0000034"

name="RNA_Polymerase" />

<species compartment="c_1" id="melr_CRPCAMP_1_CRPDimer_RNA_Polymerase" metaid="metaid_0000035"

name="melr_CRPCAMP_1_CRPDimer_RNA_Polymerase" />

<species compartment="c_1" id="melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer_Melrbs2_Melr_Melibiose_RNA_Polymerase" metaid="metaid_0000036"

name="melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer_Melrbs2_Melr_Melibiose_RNA_Polymerase" />

</listOfSpecies>

<listOfReactions>

<reaction id="reaction_0000001" metaid="metaid_00000131"

name="binding of Melrb1 using Melr" reversible="true">

<listOfReactants>

<speciesReference species="melr_Melrbs1" />

<speciesReference species="MelR" />

</listOfReactants>

<listOfProducts>

<speciesReference species="melr_Melrbs1_MelR" />

</listOfProducts>

<kineticLaw>

<math xmlns="http://www.w3.org/1998/Math/MathML">

<apply>

<times/>

<ci>c_1</ci>

<apply>

<minus/>

<apply>

<times/>

<ci>kon</ci>

<ci>melr_Melrbs1</ci>

<ci> MelR </ci>

</apply>

<apply>

<times/>

<ci>koff</ci>

<ci>melr_Melrbs1_MelR </ci>

</apply>

</apply>

</apply>

</math>

<listOfParameters>

<parameter id="kon" value="1000000" units="litre_per_mole_per_second"/>

<parameter id="koff" value="10000" units="per_second"/>

</listOfParameters>

</kineticLaw>

</reaction>

<reaction id="reaction_0000002" metaid="metaid_00000132"

name="binding of Melrb2 using Melr" reversible="true">

<listOfReactants>

<speciesReference species="melr_Melrbs2" />

<speciesReference species="MelR" />
</listOfReactants>
<listOfProducts>
<speciesReference species="melr_Melrbs2_MelR" />
</listOfProducts>
<kineticLaw>
<math xmlns="http://www.w3.org/1998/Math/MathML">
<apply>
<times/>
<ci>c_1</ci>
<apply>
<minus/>
<apply>
<times/>
<ci>kon</ci>
<ci>melr_Melrbs2</ci>
<ci> MelR </ci>
</apply>
<apply>
<times/>
<ci>koff</ci>
<ci>melr_Melrbs2_MelR </ci>
</apply>
</apply>

```
                    </apply>
                </math>
                <listOfParameters>
                    <parameter id="kon" value="1000000"
units="litre_per_mole_per_second"/>
                    <parameter id="koff" value="10000"
units="per_second"/>
                </listOfParameters>
            </kineticLaw>
        </reaction>
        <reaction id="reaction_0000003" metaid="metaid_00000133"
            name="binding of Melrb11 using Melr" reversible="true">
            <listOfReactants>
                <speciesReference species="melr_Melrbs11" />
                <speciesReference species="MelR" />
            </listOfReactants>
            <listOfProducts>
                <speciesReference species="melr_Melrbs11_MelR" />
            </listOfProducts>
            <kineticLaw>
              <math xmlns="http://www.w3.org/1998/Math/MathML">
                <apply>
                  <times/>
                  <ci>c_1</ci>
                  <apply>
```

<minus/>

<apply>

<times/>

<ci>kon</ci>

<ci>melr_Melrbs11</ci>

<ci> MelR </ci>

</apply>

<apply>

<times/>

<ci>koff</ci>

<ci>melr_Melrbs11_MelR </ci>

</apply>

</apply>

</apply>

</math>

<listOfParameters>

<parameter id="kon" value="1000000" units="litre_per_mole_per_second"/>

<parameter id="koff" value="10000" units="per_second"/>

</listOfParameters>

</kineticLaw>

</reaction>

<reaction id="reaction_0000004" metaid="metaid_00000134"

name="binding of MelrbR using Melr" reversible="true">

```
<listOfReactants>
        <speciesReference species="melr_MelrbsR" />
        <speciesReference species="MelR" />
</listOfReactants>
<listOfProducts>
        <speciesReference species="melr_MelrbsR_MelR" />
</listOfProducts>
<kineticLaw>
   <math xmlns="http://www.w3.org/1998/Math/MathML">
        <apply>
           <times/>
           <ci>c_1</ci>
           <apply>
                <minus/>
                <apply>
                   <times/>
                   <ci>kon</ci>
                   <ci>melr_MelrbsR</ci>
                   <ci> MelR </ci>
                </apply>
                <apply>
                   <times/>
                   <ci>koff</ci>
                   <ci>melr_MelrbsR_MelR </ci>
```

```
                                    </apply>
                                </apply>
                            </apply>
                        </math>
                        <listOfParameters>
                            <parameter id="kon" value="1000000" units="litre_per_mole_per_second"/>
                            <parameter id="koff" value="10000" units="per_second"/>
                        </listOfParameters>
                    </kineticLaw>
                </reaction>
                <reaction id="reaction_0000005" metaid="metaid_00000135"
                    name="binding of CRP CAMP1 using CRPDimer" reversible="true">
                    <listOfReactants>
                        <speciesReference species="melr_CRPCAMP_1" />
                        <speciesReference species="CRPDimer" />
                    </listOfReactants>
                    <listOfProducts>
                        <speciesReference species="melr_CRPCAMP_1_CRPDimer" />
                    </listOfProducts>
                    <kineticLaw>
                      <math xmlns="http://www.w3.org/1998/Math/MathML">
                        <apply>
```

<times/>

<ci>c_1</ci>

<apply>

<minus/>

<apply>

<times/>

<ci>kon</ci>

<ci> melr_CRPCAMP_1</ci>

<ci> CRPDimer </ci>

</apply>

<apply>

<times/>

<ci>koff</ci>

<ci> melr_CRPCAMP_1_CRPDimer </ci>

</apply>

</apply>

</apply>

</math>

<listOfParameters>

<parameter id="kon" value="1000000" units="litre_per_mole_per_second"/>

<parameter id="koff" value="10000" units="per_second"/>

</listOfParameters>

</kineticLaw>

```
                    </reaction>
                    <reaction id="reaction_0000006" metaid="metaid_00000136"
                        name="binding of CRP CAMP2 using CRPDimer"
reversible="true">
                        <listOfReactants>
                            <speciesReference species="melr_CRPCAMP_2" />
                            <speciesReference species="CRPDimer" />
                        </listOfReactants>
                        <listOfProducts>
                            <speciesReference
species="melr_CRPCAMP_2_CRPDimer" />
                        </listOfProducts>
                        <kineticLaw>
                          <math xmlns="http://www.w3.org/1998/Math/MathML">
                            <apply>
                              <times/>
                              <ci>c_1</ci>
                              <apply>
                                <minus/>
                                <apply>
                                  <times/>
                                  <ci>kon</ci>
                                  <ci> melr_CRPCAMP_2</ci>
                                  <ci> CRPDimer </ci>
                                </apply>
```

```
                                        <apply>
                                            <times/>
                                            <ci>koff</ci>
                                            <ci> melr_CRPCAMP_2_CRPDimer </ci>
                                        </apply>
                                    </apply>
                                </apply>
                            </math>
                            <listOfParameters>
                                <parameter id="kon" value="1000000" units="litre_per_mole_per_second"/>
                                <parameter id="koff" value="10000" units="per_second"/>
                            </listOfParameters>
                        </kineticLaw>
                    </reaction>
                    <reaction id="reaction_0000007" metaid="metaid_00000137"
                        name="binding of Melrb1 using Melr_Melibiose" reversible="true">
                        <listOfReactants>
                            <speciesReference species="melr_Melrbs1" />
                            <speciesReference species="Melr_Melibiose" />
                        </listOfReactants>
                        <listOfProducts>
                            <speciesReference species="melr_Melrbs1_Melr_Melibiose"/>
```

```
</listOfProducts>
<kineticLaw>
  <math xmlns="http://www.w3.org/1998/Math/MathML">
      <apply>
        <times/>
        <ci>c_1</ci>
        <apply>
            <minus/>
            <apply>
              <times/>
              <ci>kon</ci>
              <ci>melr_Melrbs1</ci>
              <ci>Melr_Melibiose</ci>
            </apply>
            <apply>
              <times/>
              <ci>koff</ci>
              <ci>melr_Melrbs1_Melr_Melibiose</ci>
            </apply>
        </apply>
      </apply>
  </math>
  <listOfParameters>
```

<parameter id="kon" value="1000000" units="litre_per_mole_per_second"/>

<parameter id="koff" value="10000" units="per_second"/>

</listOfParameters>

</kineticLaw>

</reaction>

<reaction id="reaction_0000008" metaid="metaid_00000138"

name="binding of Melrb2 using Melr_Melibiose" reversible="true">

<listOfReactants>

<speciesReference species="melr_Melrbs2" />

<speciesReference species="Melr_Melibiose" />

</listOfReactants>

<listOfProducts>

<speciesReference species="melr_Melrbs2_Melr_Melibiose" />

</listOfProducts>

<kineticLaw>

<math xmlns="http://www.w3.org/1998/Math/MathML">

<apply>

<times/>

<ci>c_1</ci>

<apply>

<minus/>

<apply>

<times/>

<ci>kon</ci>

<ci>melr_Melrbs2</ci>

<ci>Melr_Melibiose</ci>

</apply>

<apply>

<times/>

<ci>koff</ci>

<ci>melr_Melrbs2_Melr_Melibiose</ci>

</apply>

</apply>

</apply>

</math>

<listOfParameters>

<parameter id="kon" value="1000000" units="litre_per_mole_per_second"/>

<parameter id="koff" value="10000" units="per_second"/>

</listOfParameters>

</kineticLaw>

</reaction>

<reaction id="reaction_0000009" metaid="metaid_00000139"

name="binding of Melrb11 using Melr_Melibiose" reversible="true">

<listOfReactants>

<speciesReference species="melr_Melrbs11" />

<speciesReference species="Melr_Melibiose" />

</listOfReactants>

<listOfProducts>

<speciesReference species="melr_Melrbs11_Melr_Melibiose" />

</listOfProducts>

<kineticLaw>

<math xmlns="http://www.w3.org/1998/Math/MathML">

<apply>

<times/>

<ci>c_1</ci>

<apply>

<minus/>

<apply>

<times/>

<ci>kon</ci>

<ci>melr_Melrbs11</ci>

<ci>Melr_Melibiose</ci>

</apply>

<apply>

<times/>

<ci>koff</ci>

<ci>melr_Melrbs11_Melr_Melibiose</ci>

```
                                    </apply>
                                </apply>
                            </apply>
                        </math>
                        <listOfParameters>
                            <parameter id="kon" value="1000000" units="litre_per_mole_per_second"/>
                            <parameter id="koff" value="10000" units="per_second"/>
                        </listOfParameters>
                    </kineticLaw>
                </reaction>
                <reaction id="reaction_00000014" metaid="metaid_00000145"
                    name="binding of Melrb21 using Melr_Melibiose" reversible="true">
                    <listOfReactants>
                        <speciesReference species="melr_Melrbs21" />
                        <speciesReference species="Melr_Melibiose" />
                    </listOfReactants>
                    <listOfProducts>
                        <speciesReference species="melr_Melrbs21_Melr_Melibiose" />
                    </listOfProducts>
                    <kineticLaw>
                      <math xmlns="http://www.w3.org/1998/Math/MathML">
                        <apply>
```

<times/>

<ci>c_1</ci>

<apply>

<minus/>

<apply>

<times/>

<ci>kon</ci>

<ci>melr_Melrbs21</ci>

<ci>Melr_Melibiose</ci>

</apply>

<apply>

<times/>

<ci>koff</ci>

<ci>melr_Melrbs21_Melr_Melibiose</ci>

</apply>

</apply>

</apply>

</math>

<listOfParameters>

<parameter id="kon" value="1000000" units="litre_per_mole_per_second"/>

<parameter id="koff" value="10000" units="per_second"/>

</listOfParameters>

</kineticLaw>

```
                    </reaction>
                    <reaction id="reaction_0000010" metaid="metaid_00000140"
                        name="binding of MelrbR using Melr_Melibiose"
reversible="true">
                        <listOfReactants>
                            <speciesReference species="melr_MelrbsR" />
                            <speciesReference species="Melr_Melibiose" />
                        </listOfReactants>
                        <listOfProducts>
                            <speciesReference
species="melr_MelrbsR_Melr_Melibiose" />
                        </listOfProducts>
                        <kineticLaw>
                          <math xmlns="http://www.w3.org/1998/Math/MathML">
                            <apply>
                              <times/>
                              <ci>c_1</ci>
                              <apply>
                                <minus/>
                                <apply>
                                  <times/>
                                  <ci>kon</ci>
                                  <ci>melr_MelrbsR</ci>
                                  <ci>Melr_Melibiose</ci>
                                </apply>
```

```
                                        <apply>
                                            <times/>
                                            <ci>koff</ci>
                                            <ci>melr_Melrbs1_Melr_Melibiose</ci>
                                        </apply>
                                    </apply>
                                </apply>
                            </math>
                            <listOfParameters>
                                <parameter id="kon" value="1000000"
units="litre_per_mole_per_second"/>
                                <parameter id="koff" value="10000"
units="per_second"/>
                            </listOfParameters>
                        </kineticLaw>
                    </reaction>
                    <reaction id="reaction_0000011" metaid="metaid_00000141"
                        name="Melr_Melibiose formation" reversible="true">
                        <listOfReactants>
                            <speciesReference species="MelR" />
                            <speciesReference species="Melibiose" />
                        </listOfReactants>
                        <listOfProducts>
                            <speciesReference species="Melr_Melibiose" />
                        </listOfProducts>
```

<kineticLaw>
<math xmlns="http://www.w3.org/1998/Math/MathML">
<apply>
<times/>
<ci>c_1</ci>
<apply>
<minus/>
<apply>
<times/>
<ci>kon</ci>
<ci> MelR </ci>
<ci>Melibiose</ci>
</apply>
<apply>
<times/>
<ci>koff</ci>
<ci>Melr_Melibiose</ci>
</apply>
</apply>
</apply>
</math>
<listOfParameters>
<parameter id="kon" value="1000000" units="litre_per_mole_per_second"/>

<parameter id="koff" value="10000" units="per_second"/>

</listOfParameters>

</kineticLaw>

</reaction>

<reaction id="reaction_0000012" metaid="metaid_00000142"

name="CRP Dimerization" reversible="true">

<listOfReactants>

<speciesReference species="CRP" />

<speciesReference species="CRP" />

</listOfReactants>

<listOfProducts>

<speciesReference species="CRPDimer" />

</listOfProducts>

<kineticLaw>

<math xmlns="http://www.w3.org/1998/Math/MathML">

<apply>

<times/>

<ci>c_1</ci>

<apply>

<minus/>

<apply>

<times/>

<ci>kon</ci>

<ci> CRP </ci>

<ci> CRP </ci>

</apply>

<apply>

<times/>

<ci>koff</ci>

<ci>CRPDimer</ci>

</apply>

</apply>

</apply>

</math>

<listOfParameters>

<parameter id="kon" value="1000000" units="litre_per_mole_per_second"/>

<parameter id="koff" value="10000" units="per_second"/>

</listOfParameters>

</kineticLaw>

</reaction>

<reaction id="reaction_0000013" metaid="metaid_00000144"

name="Transcription initiation of melR" reversible="true">

<listOfReactants>

<speciesReference species="melr_CRPCAMP_1_CRPDimer" />

<speciesReference species="RNA_Polymerase" />

```
                    </listOfReactants>
                    <listOfProducts>
                        <speciesReference
species="melr_CRPCAMP_1_CRPDimer_RNA_Polymerase" />
                    </listOfProducts>
                    <kineticLaw>
                      <math xmlns="http://www.w3.org/1998/Math/MathML">
                        <apply>
                          <times/>
                          <ci>c_1</ci>
                          <apply>
                            <minus/>
                            <apply>
                              <times/>
                              <ci>kon</ci>
                              <ci>melr_CRPCAMP_1_CRPDimer</ci>
                              <ci>RNA_Polymerase</ci>
                            </apply>
                            <apply>
                              <times/>
                              <ci>koff</ci>

<ci>melr_CRPCAMP_1_CRPDimer_RNA_Polymerase</ci>
                            </apply>
                          </apply>
```

```
                                    </apply>
                                </math>
                                <listOfParameters>
                                    <parameter id="kon" value="1000000"
units="litre_per_mole_per_second"/>
                                    <parameter id="koff" value="100"
units="per_second"/>
                                </listOfParameters>
                            </kineticLaw>
                        </reaction>
                        <reaction id="reaction_0000017" metaid="metaid_00000149"
                            name="Transcription initiation of melR" reversible="true">
                            <listOfReactants>
                                <speciesReference
species="melr_CRPCAMP_1_CRPDimer_RNA_Polymerase" />
                            </listOfReactants>
                            <listOfProducts>
                                <speciesReference species="RNA_Polymerase" />
                                <speciesReference species="MelR" />
                            </listOfProducts>
                            <kineticLaw>
                           <math xmlns="http://www.w3.org/1998/Math/MathML">
                            <apply>
                              <times/>
                              <ci>c_1</ci>
```

<ci>kcat</ci>

<ci>melr_CRPCAMP_1_CRPDimer_RNA_Polymerase</ci>

</apply>

</math>

<listOfParameters>

<parameter id="kcat" value="0.1" units="per_second"/>

</listOfParameters>

</kineticLaw>

</reaction>

<reaction id="reaction_0000014" metaid="metaid_00000148"

name="Binding of Melrbs11 to Melr_Melibiose and CRPCAMP_1 to CRPDimer" reversible="true">

<listOfReactants>

<speciesReference species="melr_Melrbs11_Melr_Melibiose" />

<speciesReference species="melr_CRPCAMP_1_CRPDimer" />

</listOfReactants>

<listOfProducts>

<speciesReference species="melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer" />

</listOfProducts>

</reaction>

<reaction id="reaction_0000015" metaid="metaid_00000146"

name="Binding of Melrbs11 to Melr_Melibiose and CRPCAMP_1 to CRPDimer and Melrbs21 to Melr_Melibiose" reversible="true">

<listOfReactants>

<speciesReference species="melr_Melrbs21_Melr_Melibiose" />

<speciesReference species="melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer" />

</listOfReactants>

<listOfProducts>

<speciesReference species="melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer_Melrbs2_Melr_Melibiose" />

</listOfProducts>

</reaction>

<reaction id="reaction_0000018" metaid="metaid_00000151"

name="Transcription of melR" reversible="true">

<listOfReactants>

<speciesReference species="melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer_Melrbs2_Melr_Melibiose" />

<speciesReference species="RNA_Polymerase" />

</listOfReactants>

<listOfProducts>

<speciesReference species="melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer_Melrbs2_Melr_Melibiose_RNA_Polymerase" />

</listOfProducts>

<kineticLaw>

<math xmlns="http://www.w3.org/1998/Math/MathML">

<apply>

<times/>

<ci>c_1</ci>

<apply>

<minus/>

<apply>

<times/>

<ci>kon</ci>

<ci>melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer_Melrbs2_Melr_Melibiose</ci>

<ci>RNA_Polymerase</ci>

</apply>

<apply>

<times/>

<ci>koff</ci>

<ci>melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer_Melrbs2_Melr_Melibiose_RNA_Polymerase</ci>

</apply>

</apply>

</apply>

</math>

<listOfParameters>

<parameter id="kon" value="1000000" units="litre_per_mole_per_second"/>

<parameter id="koff" value="10000" units="per_second"/>

```
                    </listOfParameters>
                </kineticLaw>
            </reaction>
            <reaction id="reaction_0000019" metaid="metaid_00000150"
                name="Transcription initiation of melR" reversible="true">
                <listOfReactants>
                    <speciesReference
species="melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer_Melrbs2_Melr_Melibio
se_RNA_Polymerase" />
                </listOfReactants>
                <listOfProducts>
                    <speciesReference species="RNA_Polymerase" />
                    <speciesReference species="MelR" />
                </listOfProducts>
                <kineticLaw>
                <math xmlns="http://www.w3.org/1998/Math/MathML">
                    <apply>
                        <times/>
                        <ci>c_1</ci>
                        <ci>kcat</ci>

<ci>melr_Melrbs11_Melr_Melibiose_CRPCAMP_1_CRPDimer_Melrbs2_Melr_Melibiose_R
NA_Polymerase</ci>
                    </apply>
                </math>
                <listOfParameters>
```

```
                        <parameter id="kcat" value="0.1" units="per_second"/>
                    </listOfParameters>
                </kineticLaw>
            </reaction>
        </listOfReactions>
    </model>
</sbml>
```