Recent Computational Approaches to Understand Gene Regulation: Mining Gene Regulation *In Silico*

I. Abnizova^{1,*}, T. Subhankulova² and W.R. Gilks³

¹MRC-BSU Robinson Way, Cambridge, UK; ²Wellcome Trust/Cancer Research UK Gurdon Institute of Cancer and Developmental Biology, Cambridge, UK and ³Leeds University, Leeds, UK

Abstract: This paper reviews recent computational approaches to the understanding of gene regulation in eukaryotes. *Cis*-regulation of gene expression by the binding of transcription factors is a critical component of cellular physiology. In eukaryotes, a number of transcription factors often work together in a combinatorial fashion to enable cells to respond to a wide spectrum of environmental and developmental signals. Integration of genome sequences and/or Chromatin Immunoprecipitation on chip data with gene-expression data has facilitated *in silico* discovery of how the combinatorics and positioning of transcription factors binding sites underlie gene activation in a variety of cellular processes.

The process of gene regulation is extremely complex and intriguing, therefore all possible points of view and related links should be carefully considered. Here we attempt to collect an inventory, not claiming it to be comprehensive and complete, of related computational biological topics covering gene regulation, which may enlighten the process, and briefly review what is currently occurring in these areas.

We will consider the following computational areas:

- gene regulatory network construction;
- evolution of regulatory DNA;
- studies of its structural and statistical informational properties;
- and finally, regulatory RNA.

Received on: November 30, 2006 - Revised on: December 13, 2006 - Accepted on: December 15, 2006

Key Words: Gene expression, transcription factors, transcriptional regulation, computational methods, data integration.

1. INTRODUCTION

Transcription regulation has been responsible for organismal complexity and diversity in the course of biological evolution and adaptation, and it is determined largely by the context-dependent complicated behaviour of *cis*-regulatory elements and transcription factors (TFs) binding to them [1,2]. Initiation of transcription in higher organisms requires binding of multiple transcription factor molecules to transcription regulatory regions, such as promoters and enhancers.

Not very much is known about regulation of transcription in eukaryotes [3-6]. To start with, let us try to make an *inventory of known* transcription factors and their binding sites, also called *cis*-regulatory elements. It is estimated that there are around 2000 TFs in the mammalian genomes [7,8] and around 1000 in the flies and worms [9]. However, only for a minority of the TFs any known information on binding sites or interacting protein partners is available currently [10]. DNA-binding site models are available for around 500 vertebrate TFs, and for approximately 3000 genes we have the information about approximately 5000 binding sites. It means [5] that the total number of such sites in the higher organism genomes could be at least an order of thousands or

more. Thus, our knowledge of the TFs, and especially their binding sites and regulated genes, is severely limited at this present time.

Nowadays, we are given a number of full genome sequences and advancements in high-throughput technologies. Thus, *in silico* methods should be used to integrate diverse data sources toward unravelling the sophisticated and complex nature of transcriptional regulation.

From a computational point of view, one may try to understand transcription regulation *directly*, by inferring Genetic Regulatory Networks (GRN) from gene and protein interactions. We will describe GRN construction briefly in the Section 1.

Indirect approaches to understand gene regulation will include the studies of:

- (i) regulatory DNA evolution;
- (ii) statistical, information and structural properties of regulatory regions as well as
- (iii) studying of regulatory RNA. Sections 2,3 and 4 are dedicated to reviewing what is done in the corresponding areas.

2. GENE REGULATORY NETWORKS AND THEIR BUILDING BLOCKS

2.1. Overview of Gene Regulatory Networks

By definition, a **gene regulatory network** (also called a *GRN*) is a collection of DNA segments in a cell which inter-

*Address correspondence to this author at the Biostatistics Unit, MRC, Institute of Public Health, Robinson Way, Cambridge, CB2 0SR, UK; E-mail: irina.abnizova@mrc-bsu.cam.ac.uk

act with each other and with other substances in the cell, thereby governing the rates at which genes in the network are transcribed into mRNA. In GRN (for a detailed review see [11,12]), genes are nodes in a complex network, with inputs being proteins such as transcription factors, and outputs being the level of gene expression. The gene networks are only beginning to be understood, and it is a next step for biology to attempt to deduce the functions for each gene "node", and to assist in modelling of cell behaviour.

Mathematical models of GRNs have been developed to allow predictions of the models to be tested. Various modelling techniques have been used, including Boolean networks [13-15], Petri nets [16,17], Bayesian networks [18-21], graphical Gaussian models [22-24] and sets of differential equations [25,26].

Some recent advances in GRN construction and applications are well presented in a number of excellent reviews: [27] for yeast and fly, [28] for sea urchin development, [29] for comparative genomic applications. There is an elegant approach to study GRNs statistically [30] by computing significant regulatory network motifs.

2.2. Gene Networks Construction in Eukaryotes by *In Silico* Methods

Understanding and constructing of complex gene regulatory networks in higher organisms is an extremely difficult task. In contrast to prokaryotes, where transcriptional regulation can be understood in terms of induction by single factors, the regulation in eukaryotes is mainly carried out by sophisticated interactions of multiple transcription factors. Additionally, the regulatory sites are distributed over large regions of the genome including intronic sequences [31,32]. It is rare that individual binding sites are strongly conserved, only the combinatorial action gives rise to a specific control. Thus, computational methods should be used to infer gene regulation and to integrate data for GRN construction.

GRNs are Constructed Based on Different Diverse Sources of Information:

a) Transcription Factor Binding Information

The signals that determine activation and repression of specific genes in response to appropriate stimuli are one of the most important, but least understood, types of information encoded in genomic DNA. These signals, the Transcription Factor Binding Sites (TFBS), constitute an important building block of GRN. Therefore a vast number of TFBS motif search algorithms are developed. The detailed review of motif discovery methods is within the scope of this paper, so we would recommend reading the following reviews on the subject: [27, 33-42]. Some reviews focus on specific techniques such as phylogenetic footprinting [43], or on specific genomes [44]. The comparison of alternative approaches to motif discovery is discussed in the excellent works [45,46].

b) Using Information from Micro-Array Data

To improve TFBS predictions and GRN construction [47], groups of *co-regulated genes inferred from expression profiling* [48,49] are considered as a search field for putative

TFBS motifs. In [50] they build an algorithm to search for combinations of transcription factor binding sites that are enriched in a set of potentially co-regulated genes with respect to the whole genome. There was proposed 'motif regressor' algorithm [51] for discovering sequence motifs upstream of genes that undergo expression changes in a given condition. In this work the authors integrated motif discovery and genome-wide expression analysis.

c) Comparative Genomics

Comparative genomics is currently a powerful tool for searching for *phylogenetically conserved* regions and elements [52-56]. These elements then submitted to GRN construction.

In the [57] an approach to present a comparative analysis of the human, mouse, rat and dog genomes to create a systematic catalogue of common regulatory motifs in promoters and 3' untranslated regions (3' UTRs) is done. Recently, phylogenetic shadowing approach [58,59] was developed and applied to compute and statistically evaluate conservation profiles of multiple sequence alignments from closely related species. This statistical method permitted the accurate prediction transcriptional regulatory elements in human-primate comparisons, and validated the use of comparative genomic approach for deciphering primate-specific functional DNA sequences.

d) Genome Organisation: cis-Regulatory Modules Prediction

Several groups [6, 60-63] have proposed that dense clusters of motifs may diagnose regulatory regions more accurately. An approach based on *e-clustering of binding sites* [63] within *cis-regulatory modules* (CRMs) is studied in [64]. Searching for co-occurrence of binding sites that form regulatory modules has been shown to be an effective approach to increasing prediction specificity without losing sensitivity [65]. The new method [66] relying on the principle that CRMs generally contain several phylogenetically conserved binding sites for a few different TFs, allows the prediction of a vast amount of CRMs within the human genome. A subset of these is shown to be bound *in vivo* by TFs using ChIP (Chromatin Immunoprecipitation) on chip experiments (the *ChIP-on-chip* technology is briefly introduced later in the subsection 2.2 i). Their analysis reveals, among other things, that CRM density varies widely across the genome, with CRM-rich regions often being located near genes encoding transcription factors involved in development. Predicted CRMs show a surprising enrichment near the 3' end of genes and in regions far from genes.

e) Combinatorial Interaction of Transcription Factors

Many cis-regulatory elements occur in *combinatorial manner* [67-69]. The TFs that act synergistically are likely to have their *cis*-elements co-localised on the genome at specific distances apart. Therefore a number of methods use this combinatorial information to improve the TFBS prediction accuracy [70,71]. Some methods integrate combinatorial information with other data sources: thus, in [73] a computational search for functional transcription-factor (TF) combinations using phylogenetically conserved sequences and microarray-based expression data is developed.

f) Context Dependent Behaviour of cis-Regulatory Elements

Some methods utilise a *context dependent behaviour of cis-regulatory elements* [74,75]. Since sequence-dependent DNA structure may play a role in protein-DNA interactions, the exploration of sequence-based hypotheses about the composition of binding sites and the ordering of features in a regulatory region should be considered.

g) Integrating of Several Available Tools for Transcription Regulation Studies

Currently, the most advanced methods tend to integrate multiple computational and experimental information, such as [72,76,77]. In [77] the authors integrate ChIP-on-chip, motif and microarray data. With the seqVISTA [76] tool it is possible to integrate a bunch of gene regulation analysis software (including eight web servers) that is scattered over the Internet.

h) Metabolic Pathways

Metabolic biological pathway is a series of chemical reactions occurring within a cell, catalysed by enzymes. Often, the initiation of another metabolic pathway occurs as a result. Network representations of *biological pathways* [78] offer a functional view of molecular biology that is different from and complementary to sequence, expression, and structure databases. There is currently available a wide range of digital collections of pathway data, differing in organisms included, functional areas covered, details of modelling and support for dynamic pathway construction. Databases [79] that represent pathway data at the level of individual interactions make it possible to combine data from different pathways and to query by network connectivity.

i) ChIP- on- Chip Experiment Information Usage

The results of *ChIP* (Chromatin Immunoprecipitation) *on chip experiments* are utilised in [80-82]. When recently the genome-wide binding analysis like ChIP-on-chip experiments have appeared, it significantly increased the chance for more reliable identification of actual binding site regions [83,81]. The short description of the Chip-on-chip technology can be found in [84,85].

Several motif finding and network discovery algorithms are developed based on this new type of data. MDscan was designed for motif finding from sequences obtained from a ChIP-on-chip experiment, using a semi-Bayesian scoring function [86]. The improved version of MDscan, Motif Regressor [87], identifies a set of non-redundant candidate motifs using MDscan following by linear regression analysis. Then the motifs whose promoter matching scores are significantly correlated with ChIP-on-chip enrichment or downstream gene expression values are selected. However, the Motif Regressor is not widely used due to its computational intensity.

The *computational modelling of genome binding* events based on Chip-on-chip data has been extensively developed in [84,81,82,88-90]. Recently, a new method [84] was presented, Joint Binding Deconvolution. This probabilistic method combines additional experimental data about Chromatin Immunoprecipitation and sequence information to

improve the spatial resolution of the transcription factor binding locations.

Several methods [81,90,91,77] try to combine different kinds of data: [81,90] use the GRAM (Genetic Regulatory Modules) algorithm [81] to cluster genes based on both ChIP-data and gene expression data. Another recently developed computational approach, 'ReMoDiscovery' [77], exploits in a concurrent way three independent data sources: ChIP-chip data, motif information and gene expression profiles.

3. EVOLUTION STUDIES OF REGULATORY DNA SEQUENCES

3.1. Evolution of cis-Regulatory Sequences

Gene expression is central to the genotype-phenotype relationship in all organisms, and it is an important component of the genetic basis for evolutionary change in diverse aspects of phenotype. However, the evolution of transcriptional regulation remains understudied and poorly understood. No general framework exists for understanding, interpreting, and predicting how transcription evolves. The first comprehensive attempts to do so are presented in [4,92]. In [4] the detailed study of evolutionary patterns within regulatory regions and elements in terms of macro and micro changes and their rates is presented.

3.2. Comparative Genomics and Types and Rates of Evolution within Conserved Non-coding Elements

In many eukaryotic genomes only a small fraction of the DNA codes for proteins, but the non-protein coding DNA contains important elements directing the development and the physiology of the organisms, like promoters, enhancers, insulators, and micro-RNA genes. It is not easy to recognise these elements and study their molecular evolution and function from the sequence alone [93-95].

In the general case where the transcription factor target sites are not known in advance, interspecific sequence comparison is now the method [92] of choice for physically identifying putative cis-regulatory modules in the intronic or intergenic DNA sequence of given animal genes. These key regulatory units [96] of the genome are evolutionarily conserved relative to flanking sequence. Thus, cis-regulatory modules can be detected computationally by interspecific comparison of the sequence surrounding the gene of interest, recognised as a block of sequence that has remained relatively similar between the two or more species.

Following this [92] suggestion, several highly conserved non-coding sequences were identified in vertebrate [97-99] genomes recently. When some of these sequences were tested *in vivo*, indeed, the majority appeared to drive tissue-specific gene expression during early development. However, several studies indicate that the system is more complicated, with regulatory regions having an underlying pattern of evolution not directly visible from simple sequence comparisons [100-102].

To overcome this problem they propose [103] an approach to the study of the *rate of evolution* of functional non-coding sequence elements (CNEs) at a macro-evolutionary scale by the means of cross-genomic comparison of two

outgroup species. They conclude that the proposed method can be used for testing hypotheses about the rate and pattern of evolution of putative cis-regulatory elements.

When speaking about evolution types, a higher eukaryotic genome, such as human, is often considered as consisting of three sequences types, each distinguished by their mode of evolution. Purifying selection is estimated to act on 2.5–5.0% of the genome [104], whereas virtually all remaining sequences are considered to have evolved neutrally and to be not functional. The third mode of evolution, positive selection of advantageous changes, is considered rare. In [104] the authors reviewed the evolutionary evidence for the majority of human-conserved DNA lying outside of the protein-coding sequence. The authors argue that within this non-coding fraction lies at least 1 Mb of functional sequence that has accumulated many beneficial nucleotide replacements, suggesting the possibility that a significant proportion of human sequence has evolved adaptively and thus has diverged by a greater extent than expected from neutral evolution. The observation of adaptive evolution pattern on *Drosophila* conserved non-coding DNA was confirmed in [105].

Nowadays, we are given enough sequences from related organisms, thus it is possible to detect **positive selection** occurring only in a portion of a gene in a subset of the sequences. In [106] they have used this comparative approach to understand the adaptations that *E. coli* has made to colonise and survive in the urinary tract.

One can compare *regulatory mechanisms between different species* by means of genome comparison. Studies of conserved non-coding sequences [107], which are putative regulatory elements, indicate that plants have far fewer CNEs per gene than mammals, suggesting that plants have less complex regulatory mechanisms.

3.3. Regulation and Phenotypic Changes

A considerable proportion of heritable human phenotypic variation (disease) is thought to result from altered gene expression. Not only single point polymorphism mutations (SNPs) within coding regions are associated with disease: there is strong experimental and association study's evidence that SNPs within regulatory regions may cause disease [108–111].

Several reviews have argued that changes in transcriptional regulation constitute a major component of the genetic basis for *phenotypic evolution* [1, 112–116].

There have been also demonstrated by combining experimental evidence and computation that the promoter regions of human genes provide a rich source of functional single nucleotide polymorphisms [117,118]. As many as 35% of promoter SNPs may be of functional significance [118]. However, their identification and evaluation is likely to be not easy [119–121].

Progress in characterising regulatory SNPs has been limited by our relative inability to discriminate between functional and non-functional SNPs. Existing bioinformatics tools such as PupaSNP Finder (<http://www.pupasnp.org/>) [122] and rSNP_Guide <http://www.mgs.bionet.nsc.ru/mgs/systems/rsnp/>) [123] attempt to do this either by assessing the degree of evolutionary conservation displayed by the

variant site or by trying to estimate whether or not the polymorphic variant disrupts transcription factor binding. Both the TRANSFAC database [124] and an annotated database of human promoter SNPs [125,126] have usually been used to assess the potential effect of a given polymorphic variant on transcription factor binding and gene expression.

There are, however, currently no computational tools which can be used to assess directly from promoter DNA sequence whether or not a given variant is likely to alter gene expression and hence be of functional significance. In [127] an attempt to define some of the characteristics of promoter polymorphisms with functional effects on gene expression has been made. Sequence variants that altered gene expression by 1.5-fold or more were strongly biased toward a location in the core and proximal promoter regions. A recent [128] study aimed at an approach to identify DNA sequence features that could allow *in silico* estimation of the likely functional consequences of single nucleotide changes in human gene promoter regions.

3.4. Regulation and Duplication

The idea that gene duplication has a major role in evolution was developed by Susumu Ohno in his classical book “Evolution by gene duplication” (1970) and is now widely accepted as a major evolutionary force [129].

Gene duplication occurs when an error in homologous recombination, a retrotransposition event, or duplication of an entire chromosome leads to the duplication of a region of DNA containing a gene [130]. As a result, one copy of the duplicate set of genes is often freed from selective pressure. The duplicate gene may either acquire mutations that lead to a gene with a novel function or acquire deleterious mutations and become a pseudogene.

It is known [130,131] that regulatory elements participate in organism evolution. Thus, after gene duplication these elements affect sub-functionalisation and neo-functionalisation processes. Gene duplication also contributes to the evolution of gene regulatory networks.

Comparative genome analyses reveal a surprising constancy in genetic content: vertebrate genomes have only about twice the number of genes that invertebrate genomes have, and the increase is primarily due to the duplication of existing genes rather than the invention of new ones. It is suggested [132] that organismal complexity arises from progressively more elaborate regulation of gene expression. Thus, in [133] the authors propose a model whereby gene duplication and the evolution of *cis*-regulatory elements can be considered as responsible for morphological diversity and the emergence of the modern vertebrate body plan.

4. INFORMATION, STATISTICAL AND STRUCTURAL PROPERTIES OF REGULATORY REGIONS

4.1. Periodicities, Repeats and Information Content within Regulatory DNA

Regulatory Motifs are Often Placed Periodically within Regulatory DNA, the Periods Reflecting DNA 3D Structure

Multiple binding motifs and even multiple binding sites for the same motif presented in the regulatory regions are

often described as ‘regulatory clusters’ [6,60,61,133-135]. Statistical models, based on motif clustering, are helpful for finding novel CRMs in the genome, but very often they consider only site density (cluster significance) [68] and relative site affinity (such as a weighted matrix score) [135]. However, it is known that specific arrangements of binding motifs within the regulatory regions are necessary to achieve proper biological function. The biological reasons leading to a specific arrangement of sites in promoters are clear: the transcription factors, bound to promoter DNA, are also involved in specific protein–protein interactions [136-137]; therefore, the binding motifs must be distributed in the promoter in a **non-random fashion**. In other words, the arrangement of binding motifs can control the formation of 3D protein complexes involved in initiation of specific transcription.

One of the Most Well Known Periods Within Regulatory DNA is 10-11 bp

Specific arrangements between binding sites are known from many examples in biology [138,139]. *In vitro* analysis of binding site arrangements in the rat collagenase-3 promoter [140] has revealed that a 10 bp (‘helical’) phasing in binding site distribution provides maximal transcriptional activity. The importance of the ‘helical phasing’ and specific binding site arrangement was also demonstrated *in vivo* for the murine CD4 promoter [141]. The ‘helical phasing’ (10 bp) has also been demonstrated computationally [142] using a large number of proximal eukaryotic promoters [143] and the list of binding motifs available from the TRANSFAC database [9].

In [144] they explored distance preferences in the arrangement of binding motifs for five transcription factors (Bicoid, Krüppel, Hunchback, Knirps and Caudal) in a large set of *Drosophila* cis-regulatory modules (CRMs). Analysis of non-overlapping binding motifs revealed the presence of periodic signals specific to particular combinations of binding motifs. The most striking periodic signals (10 bp for Bicoid and 11 bp for Hunchback) suggest preferential positioning of some binding site combinations on the same side of the DNA helix. The authors also analysed distance preferences in arrangements of highly correlated overlapping binding motifs, such as Bicoid and Krüppel.

Genomic DNA sequences contain a wealth of information about the bendability and curvature of the DNA molecule. For example, the well-known 10-11 bp periodicities within genomes can be attributed to supercoiled structures or wrapping around nucleosomes. In [145] study, the authors analyse correlation functions of relatively long motifs such as tetramers or poly(A) sequences. Periodically placed motifs may indicate **regular protein binding** or curvature signals. The authors detected various periodic signals e.g. strong 10-11 bp oscillations of periodically placed poly(A), poly(T) or poly(W) stretches.

Composite Elements

The very successful side effect of site arrangement studies is the concept of composite elements (CEs) [146,147]. In the simplest case, a CE corresponds to a pair of individual binding motifs located at a particular distance and involved in formation of specific tertiary (DNA–protein–protein–DNA) complexes. Identical CEs may perform related func-

tions in different genes. Further development of this concept resulted in construction of a dedicated database TRANS Compel [148], combining sequences for 256 CEs from different organisms. Currently, the CE concept is widely used for finding co-localised, synergistic (antagonistic) binding motif pairs [149,150] or combinatorial arrays of motifs responsible for the formation of similar gene expression profiles [151-153].

Tandem Repeats and Regulatory Elements

Genomic sequences are highly redundant and contain many types of repetitive DNA. Fuzzy tandem repeats (FTRs) are of particular interest. They are found in regulatory regions of eukaryotic genes and are reported to interact with transcription factors.

Using TandemSWAN tool [154] have compared the structure and the occurrence of FTRs with short period length (up to 24 bp) in coding and non-coding regions including UTRs, heterochromatic, intergenic and enhancer sequences of *Drosophila melanogaster* and *Drosophila pseudoobscura*. Tandems with period three and its multiples were found in coding segments, whereas FTRs with periods multiple of six are overrepresented in all non-coding segment. Periods equal to 5–7 and 11–14 were characteristic of the enhancer regions and other non-coding regions close to genes.

Information Content

Information content (IC, also called relative entropy, see [155]) commonly used to measure a TFBS motif affinity and significance [155-157]. Statistical methods for computing the *P*-value of IC have been defined in [158,159]. Several algorithms also use the IC measure to identify optimal motifs from input sequences [160-163].

4.2. Statistics of Regulatory DNA

We unavoidably use statistics to describe and model regulatory DNA and its activity. From biological observations and computational studies [92,4] we know about some statistical properties of regulatory sequences: binding motif over-representation, motif spatial distribution, combinatorial nature and context –dependent behaviour of regulatory elements. It is also known [4] that these properties probably originated from the evolutionary processes governing the development of transcription.

Generally, when we do not have clear understanding of some process (namely, transcription regulation) we attempt to describe it statistically and learn from known examples. In this subsection we briefly review computational approaches to model regulatory DNA sequence and capture its statistical properties. These statistical models and properties are utilised further for different purposes: motif discovery, regulatory modules search, constructing evolutionary models and GRNs, etc. Some of approaches were already mentioned in the previous subsections.

A lot of statistics are developed to reflect different biological aspects of TFBS motifs: their internal position dependencies [164], their spatial distribution within regulatory regions [165,166] and genome-wide [6,61,133].

4.2.1. Models of TF Binding Motifs and their Statistics

Genomic regulatory elements are commonly represented by DNA motifs. Recurrent motifs in a collection of DNA sites are most frequently modelled by sequence patterns (also called regular expressions), or by position weight matrices (PWMs, also called profiles and position-specific scoring matrices, PSSMs). The DNA sequence patterns are simpler in representation, and have an advantage in computing their significance in the genome using statistical methods. The PWMs are able to capture information on the variability of a collection of DNA sites in a quantitative manner, which is not possible with the DNA patterns. The most important practical application of these models is their usage in scanning DNA sequences for new regulatory element candidates. Currently, there are two main databases containing information on TFs binding site profiles [9,167]. JASPAR [167] contains a smaller set that is non-redundant, while TRANSFAC [9] contains multiple profile models for some TFs.

Biological Issues Relating to the Predictions of Functional TF-Binding Sites

Site predictions with the PWMs and DNA patterns can suffer from high false positive rates if motifs are *degenerate*. Although there are limitations, the PWM models work fairly well in representing the specificity of DNA-binding sites and predicting TF-binding probability to a given site [168,169]. Some TFs are by nature moderately or poorly specific in their DNA binding and achieve higher specificity only in the *context* of other binding partners [170,171]. One also has to remember that the *chromatin structure* and DNA *methylation* play important roles in gene regulation [172-174].

Positional Dependencies Within TFBS

The simple PWM approach assumes independent contributions from each position within a DNA site towards the binding free energy of the TF (mononucleotide model). The score of an aligned substring is the log-likelihood of the substring under a product multinomial distribution. PWM scores can also be described in a physical framework as the sum of binding energies for all nucleotides aligned with the PWM [175,176].

This has been shown not to hold true in many (around 25%) situations [164,177,178]. In such cases, higher-order models are better representations and give more accurate predictions when searching for candidate sites [164,178,179,181]. Most of the higher-order models [164,178-183] assume di-nucleotide interactions, which requires less information [9] in comparison with higher orders, and has been proven by experimental data [177,178].

Many different position-dependent extensions to the basic PWMs have been proposed in the literature [164,184,185]. Some authors [186] model probabilistic motifs as n 'th order Markov chains, and furthermore, a variable-length Markov model (VLMM) [187] in case of varying dependencies within the motif is proposed.

Instead of focusing on dependencies between specific nucleotides at different positions, Xing *et al.* [188] model the distribution of conserved positions or blocks of motif positions [170] within a motif with a Markov chain.

4.2.2. CRM Statistics

Different *statistics* are typically measured within cis-regulatory modules:

- *significant maximal distance* between single motifs [189-193]. In some works [63,194,195] they model intermotif distance with Poisson distribution. A distance between motifs is expected to be constant (conserved) within regulatory modules [133,196];
- *combination of single motif occurrences* [197-199, 133] and their inter-motif *distances* [68,190,194,75, 200]. Multiple *motif occurrences* [201-202,191,51,63] when presence of multiple TFBS and their interactions are assumed;
- *density* of motifs in a window [68].

They apply a logistic regression model to assess the motif combination significance

- *dictionary* of statistically significant 'word' [203,204];
- *information content* of known motif representations [155-159].

Because regulatory regions may consist of different CRMs, they compute

- statistically significant *combinations of CRMs* [6,61]. *Distances* between modules may vary significantly [6,194]

4.2.3. Genome-wide Statistics

- *genome-level over-representation* of a motif. They compute genome-wide motif frequency as a background expectation [64,205-208], and look for genes around which they observe more motifs than expected;
- over-representation of motif combinations genome-wide [209].

4.2.4. Correspondence with Experimental Data

- correspondence between *micro-array data* and *combination of single motifs and modules* [210-214]. Sometime this correspondence has a form of joint likelihood [213];
- correspondence between motif over-representation and cross-species conservation [215].

4.2.5. Statistical Motif Discovery Algorithms

Usually, the statistical findings mentioned are utilised in statistical motif discovery algorithms. These algorithms are typically based either on *iterative refinement* or *stochastic optimisation*. Expectation maximisation (EM) [216-220] is the most widely used iterative refinement method, but variable EM [194] has also been used. The stochastic optimisation technique most commonly used for motif discovery is Gibbs sampling [221-225], sometimes combined with general Metropolis-Hastings [222,226,227]. Recently, genetic algorithms [228], evolutionary Monte Carlo [226] and simulated annealing [229,211,215] has also been developed and successfully applied by bioinformatic community.

4.3. Regulation and Structural DNA Properties

The use of **structural information** [230-236] has a lot of potential for assisting with regulatory region prediction. For example, methods to predict areas of helix destabilisation are developed to help in reducing false positives [237] and distance correlations between large sets of elements that have been used to identify over-represented correlations without the need for training [238].

The three-dimensional structure of DNA, densely packed as chromatin, inhibits transcriptional initiation *in vivo* [239]. The bendability of a region, as well as its position in DNA loops, may indicate whether it contains regulatory elements or not. Several works take DNA structure into consideration, such as [75] and [240,241] where helical parameter features are studied.

However, structural approaches have been limited in their general use so far, since derivation of the quantitative structural parameters is dependent on the small number of solved protein-DNA complex structures. The distance between the transcription start site (TSS) and translation start site (TLS) has been explicitly utilised in promoter prediction algorithms [242]. Some algorithms informally incorporate this information by restricting their search areas using the TLS as a reference [243] while others incorporate probability distributions.

One also has to remember that the **chromatin structure and DNA methylation** play important roles in gene regulation [172-174]. Large portions of the chromosomal DNA are sequestered by histones forming part of the nucleosomal structure, and are therefore not accessible for binding by the TFs. DNA methylation can inhibit interaction of the regulatory proteins with cognate DNA sites and also influence the chromatin structure. While doing genome-wide searches for putative binding sites using a motif model, one typically does not know the chromosomal regions that are open for the regulatory proteins to bind with, or the binding partners for a given TF. A blind search of the entire genome without such information usually returns a large number of sites, many of which would probably bind to the TF if the DNA sequences were open for binding, but are biologically non-functional *in vivo*. Regulatory sequences may often be outside of the nucleosome structure, or at least available a part of the time due to chromatin remodelling [244]. So the rate of non-functional site predictions in these sequences is likely to be lower than other parts of the genome. However, without the information on DNA availability, methylation status or other binding partners, the binding site predictions with individual models are likely to produce a lot of false positives.

The genetic context of a regulatory element is important for its activity, and presence of CpG-islands may be relevant factors. Both high GC content and presence of CpG-islands may indicate [245] that a region contains regulatory elements.

5. REGULATORY RNA

Not only cis-regulatory elements participate in regulation of gene transcription. It is now evident [246-250] that non-protein coding RNA (ncRNA) plays a critical role in regulating the timing and rate of protein translation.

Large scale cDNA sequencing and genome tiling array studies have shown [248] that around 50% of genomic DNA in humans is transcribed, of which 2% is translated into proteins and the remaining 98% is non-coding RNAs (ncRNAs). Today, we know **very little** about **the regulatory mechanisms and functions of these ncRNAs**.

Indeed, recent evidence [250] suggests that the majority of the genomes of mammals and other complex organisms is in fact transcribed into ncRNAs, many of which are alternatively spliced and processed into smaller products. These ncRNAs include microRNAs and snoRNAs (many if not most of which remain to be identified), as well as likely other classes of yet-to-be-discovered small regulatory RNAs, and tens of thousands of longer transcripts (including complex patterns of interlacing and overlapping sense and antisense transcripts), most of whose functions are unknown. These RNAs (including those derived from introns) appear to comprise a hidden layer of internal signals that control various levels of gene expression in physiology and development, including chromatin architecture and epigenetic memory, transcription, RNA splicing, editing, translation and turnover. Regulatory RNA may play a significant role in disease occurrence and to be responsible for a large amount of genetic variation within and between species.

The potential importance of ncRNAs is suggested by the observation that the complexity of an organism is poorly correlated with its number of protein coding genes, yet highly correlated with its number of ncRNA genes, and that in the human genome only a small fraction (2-3%) of genetic transcripts are actually translated into proteins. In the review [246] several examples of known RNA mechanisms for the regulation of protein synthesis are discussed. In eukaryotes ncRNAs have been shown [247] to operate on virtually every level of transmission of genetic information.

Recently, in [57] by means of cross-genome comparison regulatory motifs for microRNA gene are discovered. In this work a comparative analysis of the human, mouse, rat and dog genomes is done to create a systematic catalogue of common regulatory motifs in promoters and 3' untranslated regions (3' UTRs). Nearly one-half of 3'-UTR motifs, which are involved in post-transcriptional regulation, are associated with microRNAs (miRNAs), leading to the discovery of many new miRNA genes and their likely target genes. Their results suggest that previous estimates of the number of human miRNA genes were low, and that miRNAs regulate at least 20% of human genes.

6. CONCLUSIONS

Significant progress in transcription regulation understanding *in silico* has been achieved, however, much more research is required.

The field of gene expression regulation brings together researchers from several disciplines, in particular from biology, statistics and informatics. Additionally, research in the field is fairly recent and moving at a fast pace. This has resulted in a broad range of computational methods that are described with different vocabulary and different foci, different ways regarding the integration of experimental information. It is not easy to navigate within and understand this information.

Clearly, significant advances have been made over the past two and half decades in methods of the DNA regulatory elements representation, modelling and identification. However, our knowledge of the transcriptional regulation of gene expression, its variation and evolution is still limited.

It is hoped that integration of techniques and experiences across from existing approaches will give rise to refined and advanced methods with higher level of regulatory mechanism understanding than what we have seen so far.

ACKNOWLEDGEMENTS

We are thankful to Graham Ellis for support and to Rene te Boekhorst for hard work with references.

REFERENCES

- [1] Davidson, E. Genomic Regulatory Systems. *Academic Press* **2001**.
- [2] Arnone, M., Davidson, E.H. The hardwiring of development: organization and function of genomic regulatory system. *Development* **1997**, *124*: 1851-64.
- [3] Dermitzakis, M., Reymond, A., Antonarakis, S. Conserved non-genic sequences-an unexpected feature of mammalian genomes. *Nat. Rev. Genet.* **2005**, *6*: 151-157.
- [4] Wray, G.A., Hahn, M., Abouheif, E., Balhoff, J., Pizer, M., Rockman, M. and Romano, L. The Evolution of Transcriptional Regulation in Eukaryotes. *Mol. Biol. Evol.* **2003** *20*(9): 1377-1419.
- [5] Yuh, C., Bolouri, H., Davidson, E. Genomic cis-regulatory logic: functional analysis and computational model of a sea urchin gene control system. *Science* **1998**, *279*: 1896-1902.
- [6] Berman, B., Nibu, Y., Pfeiffer, P., Tomancak, P., Celniker, S., Rubin, G., Levine, M., Eisen, M. Exploiting TFBS clustering to identify CRM involved in pattern formation in *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* **2002**, *99*(2): 757-762.
- [7] Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaanty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyas, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigo, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W.J., Kirby, A., Kolbe, D.L., Korf, I., Kucherlapati, R.S., Kulbokas, E.J., Kulp, D., Landers, T., Leger, J.P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D.R., Mardis, E.R., Matthews, L., Mauceli, E., Mayer, J.H., McCarthy, M., McCombie, W.R., McLaren, S., McLay, K., McPherson, J.D., Meldrum, J., Meredith, B., Mesirov, J.P., Miller, W., Miner, T.L., Mongin, E., Montgomery, K.T., Morgan, M., Mott, R., Mullikin, J.C., Muzny, D.M., Nash, W.E., Nelson, J.O., Nhan, M.N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M.J., Okazaki, Y., Oliver, K., Larty, E.O., Pachter, L., Parra, G., Pepin, K.H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C.S., Poliakov, A., Ponce, T.C., Ponting, C.P., Potter, S., Quail, M., Reymond, A., Roe, B.A., Roskin, K.M., Rubin, E.M., Rust, A.G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M.S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J.B., Slater, G., Smit, A., Smith, D.R., Spencer, B., Stabenau, A., Strange-Thomann, N.S., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Vidal, A.U., Vinson, J.P., von, Niederhauser, A.C., Wade, C.M., Wall, M., Weber, R.J., Weiss, R.B., Wendl, M.C., West, A.P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R.K., Winter, E., Worley, K.C., Wyman, D., Yang, S., Yang, S.P., Zdobnov, E.M., Zody, M.C., Lander, E.S., Initial sequencing and comparative analysis of the mouse genome. *Nature* **2002**, *420*: 520-562.
- [8] Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczy, J., LeVine, R., McEwan, P., McKernan, K., Meldrum, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissole, S.L., Wendl, M.C., Delehaanty, K.D., Miner, T.L., Delehaanty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brothier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickinson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la, Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Honkamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*: 860-921.
- [9] Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **2003**, *31*: 374-378.
- [10] Levine, M., Tjian, R. Transcription regulation and animal diversity. *Nature* **2003**, *424*: 147-151.
- [11] Bower, J. and Bolouri, H. (editors). Computational Modeling of Genetic and Biochemical Networks, Computational Molecular Biology Series, *MIT Press* **2001**.
- [12] Smolen, P., Baxter, D.A., Byrne, J.H. Modeling transcriptional control in gene networks - methods, recent results, and future directions. *Bull. Math. Biol.* **2000**, *62*: 247-92.
- [13] Shmulevich, I.E., Dougherty, R., Kim, S., Zhang, W. Probabilistic Boolean Networks. A Rule-based Uncertainty Model for Gene Regulatory Networks. *Bioinformatics* **2002**, *18* (2): 261-274.

- [14] Brazma, A. and Schlitt, T. Reverse engineering of gene regulatory networks: a finite state linear model, *Genome Biol.* **2003**, 4: P5.
- [15] Akutsu, T., Miyano, S., Kuhara, S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.* **1999**, 17-28.
- [16] Chen, M. Modelling and Simulation of Metabolic Networks: Petri Nets Approach and Perspective. In: *Modelling and Simulation 2002*: Proceedings of ESM2002, Amorski K. (ed.), Darmstadt, 441-444.
- [17] Genrich, H., Kueffner, R. and Voss, K. Executable Petri Net Models for the Analysis of Metabolic Pathways. *Int. J. Soft. Tools Technol. Tran.* **2001**, 3: 394-404.
- [18] Friedman, N., Goldszmidt, M. and Wyner, A. Data Analysis with Bayesian Networks: A Bootstrap Approach, *UAI 99* **1999**.
- [19] Barash, Y. and Friedman, N. Context-Specific Bayesian Clustering for Gene Expression Data. *J. Comput. Biol.* **2002**, 9(2): 169-91.
- [20] Zou, M. and Conzen, S. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course micro-array data. *Bioinformatics* **2005**, 21(1): 71 - 79.
- [21] Hartemink, A. Bayesian Networks and Informative Priors: Transcriptional Regulatory Network Models. In *Bayesian Inference for Gene Expression and Proteomics* **2006**, Do, K.-A., Müller, P., & Vannucci, M., eds. Cambridge University Press Cambridge, UK, 401-424.
- [22] Toh, H. and Horimoto, K. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics* **2002**, 18: 287-297.
- [23] Wu, X., Ye, Y. and Subramanian, K. Interactive analysis of gene interactions using graphical Gaussian model. ACM SIGKDD Workshop on Data Mining in *Bioinformatics* **2003**, 3: 63-69.
- [24] Waddell, P.J. and Kishino, H. Cluster inferences methods and graphical models evaluated on NCI60 microarray gene expression data. *Genome Inform.* **2000**, 11: 129-140.
- [25] Chen, T., He, H.L., Church, G.M. Modeling gene expression with differential equations. *Pac. Symp. Biocomput.* **1999**, 29-40.
- [26] Hasty, J., McMillen, D., Isaacs, F. and Collins, J. Computational studies of gene regulatory networks: *in numero* molecular biology. *Nat. Rev. Genet.* **2001**, 2(6): 268-279.
- [27] Siggia E. Computational methods for transcriptional regulation, *Curr. Op. Genetics Dev.* **2005**, 15: 214-221
- [28] Oliveri, P., Davidson, E., Gene regulatory network controlling embryonic specification in the sea urchin, *Curr. Op. Genetics Dev.* **2004**, 14: 351-360
- [29] Li, H., Wong, W., Model based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **2001**, 98: 31-36.
- [30] Shen-Orr, S., Milošević, R., Mangan, S. and Alon, U. Network motifs in transcriptional regulation network of *E. coli*, *Nat. Genet.* **2002**, 31: 64-68.
- [31] Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., Gingeras, TR. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **2004**, 116: 499-509.
- [32] Euskirchen, G., Royce, T.E., Bertone, P., Martone, R., Rinn, J.L., Nelson, F.K., Sayward, F., Luscombe, N.M., Miller, P., Gerstein, M., Weissman, S., Snyder, M. CREB binds to multiple loci on human chromosome 22. *Mol. Cell Biol.* **2004**, 24: 3804-3814.
- [33] Abnizova, I. and Gilks, W.R. Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in vertebrate genomes. *Brief. Bioinform.* **2006**, 7(1): 48 - 54.
- [34] Vavouri, T. and Elgar, G. Prediction of cis-regulatory elements using binding site matrices-the success, the failures and the reasons for both. *Curr. Opin. Genet. Develop.* **2005**, 15(4): 395-402.
- [35] Brazma, A., IJonassen, I., Vilo, J., Ukkonen E. Predicting gene regulatory elements in silico on a genomic scale, *Genome Res.* **1998**, 8: 1202-1215.
- [36] Pavesi, G., Mauri, G., Pesole, G. *In silico* representation and discovery of transcription factor binding sites. *Brief Bioinform.* **2004**, 5(3): 217-36.
- [37] Wasserman, W.W., Krivan, W. In silico identification of metazoan transcriptional regulatory regions. *Naturwissenschaften* **2003**, 90(4): 156-66.
- [38] Bulky, M.L. Computational prediction of transcription-factor binding site locations. *Genome Biol.* **2003**, 5: 201.
- [39] Sandve, G.K. and Finn, D. A survey of motif discovery methods in an integrated framework. *Biol. Direct.* **2006**, 1: 11.
- [40] van Helden, J. Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics* **2004**, 20: 399-406.
- [41] Frith, M., Fu, Y., Chen, J., Hansen, U., Weng, Z. Detection of functional motifs via statistical representation. *Nucleic Acid Res.* **2004**, 32: 1372-1381.
- [42] Wasserman, W.W., Fickett, J.W. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **1998**, 278: 167-181.
- [43] Blanchette, M., Bataille, A., Chen, X., Poitras, C., Laganière, C., Deblois, G., Giguere, V., Feretti, V., Bergeron, D., Coulombe, B., Robert, F. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression, *Genome Res.* **2006**, 16: 656-68.
- [44] Hannehalli, S., Levy, S.: Promoter prediction in the human genome. *Bioinformatics* **2001**, 17(Suppl 1): S90-96.
- [45] Tompa, M., Li N., Bailey, T.L., Church, GM, 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. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **2005**, 23: 137-44.
- [46] Li, N. and Tompa, M. Analysis of computational approaches for motif discovery. *Algorithms Mol. Biol.* **2006**, 1: 8.
- [47] Hartemink, A.J., Gifford, D., Jaakkola, T., Young, R. Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput.* **2002**, 437-49.
- [48] Kielbasa, S.M., Blüthgen, N., Sers, C., Schäfer, R., Herzel, H. Prediction of cis-regulatory elements of coregulated genes. *Genome Inform.* **2004**, 15: 117-124.
- [49] Pilpel, Y., Sudarsanam, P., Church, G.M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.* **2001**, 29: 153-159.
- [50] Kreinman, G. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes, *Nucleic Acids Res.* **2004**, 32(9): 2889 - 2900.
- [51] Conlon, E.M., Liu, X.S., Lieb, J.D., Liu, J.S. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci. USA* **2003**, 100(6): 3339-44.
- [52] Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., Lawrence, C.E. Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.* **2000**, 26: 225-228.
- [53] Dieterich, C., Wang, H., Rateitschak, K., Luz, H., Vingron, M. CORG: a database for comparative regulatory genomics. *Nucleic Acids Res.* **2003**, 31: 55-57.
- [54] Blanchette, M., Schwikowski, B., Tompa, M. Algorithms for phylogenetic footprinting. *J. Computational Biol.* **2002**, 2: 11-23.
- [55] Dermitzakis, E.T. and Clark, A.G. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **2002**, 19: 1114-1121.
- [56] Hardison, R. Comparative genomic. *PLoS Biol.* **2003**, 1(2): e58.
- [57] Xie, X., Jun, L., Kulbokas, E., Golub, T., Mootha, W., Lindblad-Toh, K., Lander, E., Kellis, M. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **2005**, 1038: 3441.
- [58] Bofelli, D., Nobrega, M., Rubin, E. Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.* **2005**, 6: 151-157.
- [59] Ovcharenko, I., Boffelli, D., Loots, G. eShadow: a tool for comparing closely related sequences. *Genome Res.* **2004**, 14(6): 1191-8.
- [60] Markstein, M., Markstein, P., Markstein, V. and Levine, M.S. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl. Acad. Sci. USA* **2002**, 99: 763-768.
- [61] Rajewsky, N., Vergassola, M., Gaul, U. and Siggia, E.D. Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinform.* **2002**, 3: 30.
- [62] Bailey, T.L. and Noble, W.S. Searching for statistically significant regulatory modules. *Bioinformatics* **2003**, 19(Suppl 2): II16-II25.
- [63] Frith, M.C., Li, M.C., Weng, Z. Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* **2003**, 31: 3666-3668.
- [64] Rebeiz, M., Reeves, N. and Posakony, J. SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. *Proc. Natl. Acad. Sci. USA* **2002**, 99(15): 9888-9893.
- [65] Wasserman, W.W., Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **2004**, 5: 276-287.

- [66] Blanchette, M., Bataille, A.R., Chen, X., Poitras, C., Laganier, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D., Coulombe, B., Robert, F. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* **2006**, *16*: 656-68.
- [67] Michelson A. Deciphering genetic regulatory codes: a challenge for functional genomics. *Proc. Natl. Acad. Sci. USA* **2002**, *99*: 546-548.
- [68] Kel-Margoulis, O. V., Tchekmenev, D., Kel, A. E., Goessling, E., Hornischer, K., Lewicki-Potapov, B. and Wingender, E. Composition-sensitive analysis of the human genome for regulatory signals. *In Silico Biol.* **2003**, *3*: 0013.
- [69] Wagner, A. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **1999**, *15*(10): 776-84.
- [70] Hannehalli, S. and Levy, S. Predicting transcription factor synergism. *Nucleic Acids Res.* **2002**, *30*: 4278-4284.
- [71] Qiu, P., Ding, W., Jiang, Y., Greene, J.R. and Wang, L. Computational analysis of composite regulatory elements. *Mamm. Genome* **2002**, *13*: 327-332.
- [72] Blüthgen, N., Kielbasa, S., and Herzel, H. Inferring combinatorial regulation of transcription *in silico*. *Nucleic Acids Res.* **2005**, *33*(1): 272-279.
- [73] Zhou, Z., Shendure, J. and Church, J. Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res.* **2005**, *15*: 848-855.
- [74] Nguen, D. & D'haeseleer P., Deciphering principles of transcription regulation in eukaryotic genomes. *Mol. Sys. Biol.* **2006**, 0012.
- [75] Beiko, R.G., Charlebois, R.L. GANN: genetic algorithm neural networks for the detection of conserved combinations of features in DNA. *BMC Bioinform.* **2005**, *6*: 36.
- [76] Hu, Z., Fu, Y., Halees, A., Kielbasa, S. and Weng, S. SeqVISTA: a new module of integrated computational tools for studying transcriptional regulation. *Proc. Natl. Acad. Sci. USA* **2004**, *32*: W235-W241.
- [77] Lemmens, K., Dhollander, T., De Bie, T., Monsieurs, P., Engelen, K., Smets, B., Winderickx, J., De Moor, B., Marchal, K. Inferring transcriptional modules from ChIP-chip, motif and microarray data. *Genome Biol.* **2006**, *7*(5): R37.
- [78] van Helden, J., Wernisch, L., Gilbert, D. Wodak, S. Graph-based analysis of metabolic networks. *Ernst Schering Research Foundation Workshop Volume 38: Bioinformatics and Genome Analysis*. **2002**, Editors: Mewes, H., Weiss, B., Seidel, H. Springer-Verlag, ISBN 3-540-42893-3: 245-274, Berlin Heidelberg.
- [79] Schaefer, C. Pathway databases, *Ann. N.Y. Acad. Sci.* **2004**, *1020*: 77-91.
- [80] Buck, M.J. and Lieb, J. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **2004**, *83*(3): 349-60.
- [81] Bar-Joseph, Z., Gerber, G., Rinaldi, N., Yoo, J., Gordon, B., Fraenkel, E., Jaakkola, T., Young, R., Gifford, D. Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* **2003**, *21*(11): 1337-42.
- [82] Harbison, C., Gordon, B., Lee, T., Rinaldi, N., MacIsaac, K., Danford, T., Gerber, G., Hanne, N., Tagne, J., Reynolds, D., Yoo, J., Jennings, E., Pokholok, D., Zeitlinger, J., Dowell, T., Kellis, M., Rolfe, A., Takusagawa, K., Lander, E., Gifford, D., Fraenkel, E., and Young, R. Transcriptional regulatory code of a eukaryotic genome. *Nature* **2004**, *431*(7004): 99-104.
- [83] Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R.A., Dynlacht, B.D. E2F integrates cell cycle progression with DNA repair, replication, and G₂/M checkpoints. *Genes Dev.* **2002**, *16*: 245-256.
- [84] Qi, Y., Rolfe, A., MacIsaac, K., Gerber, G., Pokholok, D., Zeitlinger, J., Danford, T., Dowell, T., Fraenkel, E., Jaakkola, T., Young, R., Gifford, D. High-resolution computational models of genome binding events. *Nat. Biotechnol.* **2006**, *24*(8): 963-70.
- [85] Hong, P., Liu, S., Zhou, Q., Lu, X., Liu, J. and Wong, W. A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics* **2005**, *21*(11): 2636-43.
- [86] Liu, X.S., Brutlag, D. and Liu, J. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* **2002**, *20*(8): 835-9.
- [87] Liu, Y., Wei, L., Batzoglou, S., Brutlag, D., Liu, J. and Liu, S. A suite of web-based programs to search for transcriptional regulatory motifs. *Nucleic Acids Res.* **2004**, *32*(Web Server issue): W204-7.
- [88] Lee, T., Rinaldi, N., Roberts, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hanne, N., Harbison, C., Thompson, C., Simon, I., Zeitlinger, J., Jennings, E., Murray, H., Gordon, B., Ren, B., Wurick, J., Tagne, J., Volkert, T., Fraenkel, E., Gifford, D., Young, R. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **2002**, *298*(5594): 799-804.
- [89] Schreiber, J., Jenner, R., Murray, H., Gerber, G., Gifford, D., Young, R. Coordinated binding of NF-kappaB family members in the response of human cells to lipopolysaccharide. *Proc. Natl. Acad. Sci. USA* **2006**, *103*(15): 5899-904.
- [90] Takusagawa, K.T. and Gifford, D. Negative information for motif discovery. *Pac. Symp. Biocomput.* **2004**, 360-71.
- [91] Hartemink, A.J., Gifford, D., Jaakkola, T., Young, R. Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput.* **2002**, 437-49.
- [92] Cameron, R. and Chow, S. and Berney, K. and Chiu, T. and Yuan, Q. and Krämer, A. and Helguero, A. and Ransick, A. and Yun, M. and Davidson, E. An evolutionary constraint: Strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. *Proc. Natl. Acad. Sci. USA* **2005**, *102* (33): 11769-11774.
- [93] Dermitzakis, E.T., and Clark, A.G. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **2002**, *19*: 1114-1121.
- [94] Tautz, D. Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.* **2000**, *10*: 575-579.
- [95] Ludwig, M., Bergman, C., Patel, H. and Kreitman, M. Evidence for stabilizing selection in eukaryotic enhancer element. *Nature* **2000**, *403*: 564-567.
- [96] Britten, R. J. & Davidson, E. H. *Q. Rev. Biol.* **1971**, *46*: 111-138.
- [97] Woolfe, A., Goodson, M., Goode, D., Snell, P., Smith, S., Vavouri, T., McEwen, G., Gilks, W., Walter, K., Abnizova, I., Edwards, Y., and Elgar, G., Highly conserved non coding sequences are associated with developmental control genes in vertebrates. *PLoS Biol.* **2005**, *3*: e7.
- [98] Bofelli, D., Nobrega, M., Rubin, E. Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.* **2005**, *6*: 151-157.
- [99] Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W., Mattick, H., Haussler, D. Ultraconserved elements in human genome. *Science* **2004**, *304*: 1321.
- [100] Levy, S., Hannehalli, S., Workman, C. Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics* **2001**, *17*: 871-877.
- [101] Hancock, J. M., Shaw, P. Bonneton, F. and Dover, G. High sequence turnover in the regulatory regions of the developmental gene hunchback in insects. *Mol. Biol. Evol.* **1999**, *16*: 253-265.
- [102] Tautz, D. Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.* **2002**, *10*: 575-579.
- [103] Wagner, G. P., Fried, C., Prohaska, S. J. & Stadler, P. F. Divergence of Conserved Non-Coding Sequences: Rate Estimates and Relative Rate Tests. *Mol. Biol. Evol.* **2004**, *21*: 2116-2121.
- [104] Ponting, C. and Lunter G. Signatures of adaptive evolution within human non-coding sequence, *Hum. Mol. Genet.* **2006** *15* (Review Issue 2): R170-R175.
- [105] Andolfatto P., Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **2005**, *437*: 1149-1152.
- [106] Chen, S., Hung, C., Xu, C., Reigstad, C., Magrini, V., Sabo, A., Blasiar, A., Bieri, T., Meyer, R., Ozersky, P., Armstrong, T., Fulton, R., Latreille, J., Spieth, J., Hooton, T., Mardis, E., Hultgren, J. and Gordon J. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach. *Proc. Natl. Acad. Sci. USA* **2006**, *103*(15): 5977 - 5982.
- [107] Lockton, S., Gaut, B. Plant conserved non-coding sequences and paralogue evolution. *Trends Genet.* **2005**, *21*(1): 60-65.
- [108] Monsuur, A.J., de, Bakker, P.I., Alizadeh, B.Z., Zhernakova, A., Bevova, M.R., Strengman, E., Franke, L., van't, Slot, R., van, Belzen, M.J., Lavrijsen, I.C., Diosdado, B., Daly, M.J., Mulder, C.J., Mearin, M.L., Meijer, J.W., Meijer, G.A., van, Oort, E., Wapenaar, M.C., Koelman, B.P., Wijmenga, C. Myosin IXB variant increases the risk of celiac disease and points toward a primary intestinal barrier defect. *Nat. Genet.* **2005**, *37*: 1341-4.
- [109] Plenge, R.M., Padyukov, L., Remmers, E.F., Purcell, S., Lee, A.T., Karlson, E.W., Wolfe, F., Kastner, D.L., Alfredsson, L., Altshuler, D., Gregersen, P.K., Klareskog, L., Rioux, J.D. Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. *Am. J. Hum. Genet.* **2005**, *77*: 1044-60.
- [110] Ueda, H., Howson, J.M., Esposito, L., Heward, J., Snook, H., Chamberlain, G., Rainbow, D.B., Hunter, K.M., Smith, A.N., Di, Genova, G., Herr, M.H., Dahlman, I., Payne, F., Smyth, D., Lowe, C., Twells, R.C., Howlett, S., Healy, B., Nutland, S., Rance, H.E., Everett, V., Smink, L.J., Lam, A.C., Cordell, H.J., Walker, N.M., Bordin, C., Hulme, J., Motzo, C., Cucca, F., Hess, J.F., Metzker,

- M.L., Rogers, J., Gregory, S., Allahabadia, A., Nithiyannan, R., Tuomilehto-Wolf, E., Tuomilehto, J., Bingley, P., Gillespie, K.M., Undlien, D.E., Ronningen, K.S., Guja, C., Ionescu-Tirgoviste, C., Savage, D.A., Maxwell, A.P., Carson, D.J., Patterson, C.C., Franklyn, J.A., Clayton, D.G., Peterson, L.B., Wicker, L.S., Todd, J.A., Gough, S.C. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **2003**, 423: 506-11.
- [111] Morahan, G., Huang, D., Ymer, S.I., Cancilla, M.R., Stephen, K., Dabadghao, P., Werther, G., Tait, B.D., Harrison, L.C., Colman, P.G. *Nat. Genet.* **2001**, 27: 218-21.
- [112] Stern, D. Perspective: evolutionary developmental biology and the problem of variation. *Evolution* **2000**, 54: 1079-1091.
- [113] Tautz, D. Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.* **2000**, 10: 575-57.
- [114] Purugganan, M. D. The molecular population genetics of regulatory genes. *Mol. Ecol.* **2000**, 9: 1451-1461.
- [115] Wray, G. A., and Lowe, C. Developmental regulatory genes and echinoderm evolution. *Syst. Biol.* **2000**, 49: 28-51.
- [116] Wilkins, A. S. The evolution of developmental pathways. *Sinauer Associates* **2002**, Sunderland, Mass.
- [117] Buckland, P., Hoogendoorn, B., Coleman, S., Guy, C., Smith, S., O'Donovan, M. Strong bias in the location of functional promoter polymorphisms. *Hum. Mutat.* **2005**, 24: 35-42.
- [118] Hoogendoorn, B., Coleman, S. L., Guy, C. A., Smith, S. K., O'Donovan, M. C. and Buckland, P. R. Functional analysis of polymorphisms in the promoter regions of genes on 22q11. *Hum. Mutat.* **2004**, 24: 35-42.
- [119] Mooney, S. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief. Bioinform.* **2005**, 6: 44-56.
- [120] Pastinen, T. and Hudson, T. J. Cis-acting regulatory variation in the human genome. *Science* **2004**, 306: 647-650.
- [121] Hudson, T. J. Wanted: regulatory SNPs. *Nat. Genet.* **2003**, 33: 439-440.
- [122] Conde, L., Vaquerizas, J. M., Santoyo, J., Al-Shahrour, F., Ruiz-Llorente, S., Robledo, M. and Dopazo, J. PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res.* **2004**, 32: W242-W248.
- [123] Ponomarenko, J. V., Merkulova, T. I., Orlova, G. V., Fokin, O. N., Gorshkova, E. V., Frolov, A. S., Valuev, V. P. and Ponomarenko, M. P. rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation. *Nucleic Acids Res.* **2003**, 31: 118-121.
- [124] Kang, H. J., Choi, K. O., Kim, B. D., Kim, S. and Kim, Y. J. FESD: a Functional Element SNPs Database in human. *Nucleic Acids Res.* **2005**, 33: D518-D522.
- [125] Tahira, T., Baba, S., Higasa, K., Kukita, Y., Suzuki, Y., Sugano, S. and Hayashi, K. dbQSNP: a database of SNPs in human promoter regions with allele frequency information determined by single-strand conformation polymorphism-based methods. *Hum. Mutat.* **2005**, 26: 69-77.
- [126] Chuzhanova, N. A., Anassiss, E. J., Ball, E. V., Krawczak, M. and Cooper, D. N. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum. Mutat.* **2003**, 21: 28-44.
- [127] Khan, I., Mort, M., Buckland, P., Cooper, D., O'Donovan, M., and Chuzhanova, N. In silico discrimination of single nucleotide polymorphisms and pathological mutations in human gene promoter regions by means of local DNA sequence context and regularity. *In Silico Biol.* **2006**, 6: 0003.
- [128] Ohno, S. Evolution by gene duplication. *Springer-Verlag* **1970**.
- [129] Zhang, J. Evolution by gene duplication: an update. *Trends Ecol. Evol.* **2003**, 18(6): 292-298.
- [130] Duarte, J., Cui, L., Wall, P., Zhang, Q., Zhang, X., Leebens-Mack, J., Ma, H., Altman, N. and dePamphilis, C. Expression Pattern Shifts Following Duplication Indicative of Sub-functionalization and Neo-functionalization in Regulatory Genes of *Arabidopsis*. *Mol. Biol. Evol.* **2005**, 23(3): 469-478.
- [131] Levine, M. and Tjian R., Transcription regulation and animal diversity. *Nature* **2003**, 424: 147-151.
- [132] McEwen, G., Woolfe A., Goode D., Vavouri T., Callaway H. and Elgar G. Ancient duplicated conserved noncoding elements in vertebrates: A genomic and functional analysis. *Genome Res.* **2006**, 15: 451-465.
- [133] Wagner, A. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **1999**, 15: 776-784.
- [134] Pickert, L., Reuter, I., Klawonn, F. and Wingender, E. Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics* **1998**, 14: 244-251.
- [135] Lifanov, A.P., Makeev, V.J., Nazina, A.G. and Papatsenko, D.A. Homotypic regulatory clusters in *Drosophila*. *Genome Res.* **2003**, 13: 579-588.
- [136] Chytil, M., Peterson, B.R., Erlanson, D.A. and Verdine, G.L. The orientation of the AP-1 heterodimer on DNA strongly affects transcriptional potency. *Proc. Natl Acad. Sci. USA* **1998**, 95: 14076-14081.
- [137] Remenyi, A., Tomilin, A., Scholer, H.R., and Wilmanns, M. Differential activity by DNA-induced quaternary structures of POU transcription factors. *Biochem. Pharmacol.* **2002**, 64: 979-984.
- [138] Dion, V. and Coulombe, B. Interactions of a DNA-bound transcriptional activator with the TBP-TFIIA-TFIIB-promoter quaternary complex. *J. Biol. Chem.* **2003**, 278: 11495-11501.
- [139] Inoue, J., Sato, R. and Maeda, M. Multiple DNA elements for sterol regulatory element-binding protein and NF-Y are responsible for sterol-regulated transcription of the genes for human 3-hydroxy-3-methylglutaryl coenzyme A synthase and squalene synthase. *J. Biochem.* **1998**, 123: 1191-1198.
- [140] D'Alonzo, R.C., Selvamurugan, N., Karsenty, G. and Partridge, N.C. Physical interaction of the activator protein-1 factors c-Fos and c-Jun with Cbfa1 for collagenase-3 promoter activation. *J. Biol. Chem.* **2002**, 277: 816-822.
- [141] Sarafova, S. and Siu, G. Precise arrangement of factor-binding sites is required for murine CD4 promoter function. *Nucleic Acids Res.* **2000**, 28: 2664-2671.
- [142] Ioshikhes, I., Trifonov, E.N. and Zhang, M.Q. Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proc. Natl Acad. Sci. USA* **1999**, 96: 2891-2895.
- [143] Praz, V., Perier, R., Bonnard, C. and Bucher, P. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res.* **2002**, 30: 322-324.
- [144] Makeev, V., Lifanov, A., Nazina, N. and Papatsenko, D. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucl. Acids Res.* **2003**, 31(20): 6016-6026.
- [145] Kumar, L., Futschik, M. and Herzel, H., DNA motifs and sequence periodicities. *In Silico Biol.* **2006**, 6: 0008.
- [146] Diamond, M.I., Miner, J.N., Yoshinaga, S.K. and Yamamoto, K.R. Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. *Science* **1990**, 249: 1266-1272.
- [147] Kel, A., Kel-Margoulis, O., Ivanova, T., Wingender, E. Cluster-Scan: A Tool for Automatic Annotation of Genomic Regulatory Sequences by Searching for Composite Clusters. *Proc. German Conf. Bioinform.* **2001**, 96-101.
- [148] Kel-Margoulis, O.V., Kel, A.E., Reuter, I., Deineko, I.V. and Wingender, E. TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.* **2002**, 30: 332-334.
- [149] Hannenhalli, S. and Levy, S. Predicting transcription factor synergism. *Nucleic Acids Res.* **2002**, 30: 4278-4284.
- [150] Qiu, P., Ding, W., Jiang, Y., Greene, J.R. and Wang, L. Computational analysis of composite regulatory elements. *Mamm. Genome* **2002**, 13: 327-332.
- [151] Wasserman, W.W. and Fickett, J.W. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **1998**, 278: 167-181.
- [152] Krivan, W. and Wasserman, W.W. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* **2001**, 11: 1559-1566.
- [153] Pilpel, Y., Sudarsanam, P. and Church, G.M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.* **2001**, 29: 153-159.
- [154] Boeva, V., Regnier, M., Papatsenko, D., Makeev, V. Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics* **2006**, 22(6): 676-684.
- [155] Berg, O.G., von Hippel, P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **1987**, 193: 723-750.
- [156] Schneider, T.D., Stormo, G.D., Gold, L., Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **1986**, 188: 415-431.
- [157] Stormo, G.D., Fields, D.S. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* **1998**, 23: 109-113.
- [158] Hertz, G.Z., Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **1999**, 15: 563-577.

- [159] Nagarajan, N., Jones, N., Keich, U. Computing the *P*-value of the information content from an alignment of multiple sequences. *Bioinformatics* **2005**, *21*: i311-i318.
- [160] Stormo, G.D., Hartzell, G.W. III. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA* **1989**, *86*: 1183-1187.
- [161] Hertz, G.Z., Hartzell, G.W. III, Stormo, G.D. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* **1990**, *6*: 81-92.
- [162] Lawrence, C.E., Reilly, A.A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **1990**, *7*: 41-51.
- [163] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **1993**, *262*: 208-214.
- [164] Zhou, Q., Liu, J.S. Modelling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* **2004**, *20*: 909-916.
- [165] FitzGerald, P., Shlyakhtenko, A., Mir A. and Vinson, C. Clustering of DNA Sequences in Human Promoters. *Genome Res.* **2004**, *14*: 1562-1574.
- [166] Papatsenko, D.A., Makeev, V.J., Lifanov, A.P., Regnier, M., Nazina, A.G. and Desplan, C. Extraction of functional binding sites from unique regulatory regions: the *Drosophila* early developmental enhancers. *Genome Res.* **2002**, *12*: 470-481.
- [167] Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **2004**, *32*: D91-D94.
- [168] Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **2000**, *16*: 16-23.
- [169] Stormo, G.D., Fields, D.S. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* **1998**, *23*: 109-113.
- [170] Kechris, K.J., van Zwet, E., Bickel, P.J., Eisen, M.B. Detecting DNA regulatory motifs by incorporating positional trends in information content. *Genome Biol.* **2004**, *5*(7): R50.
- [171] Scherf, M., Klingenhoff, A., Werner, T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* **2000**, *297*(3): 599-606.
- [172] Ashraf, S.I., Ip, Y.T. Transcriptional control: repression by local chromatin modification. *Curr. Biol.* **1998**, *8*: R683-R686.
- [173] Razin, A. CpG methylation, chromatin structure and gene silencing—a three-way connection. *EMBO J.* **1998**, *17*: 4905-4908.
- [174] Farkas, G., Leibovitch, B.A., Elgin, S.C. Chromatin organization and transcriptional control of gene expression in *Drosophila*. *Gene* **2000**, *253*: 117-136.
- [175] Barash, Y., Elidan, G., Friedman, N., Caplan, T. Modeling Dependencies in Protein-DNA Binding Sites, In *Proc. of the 7th Int. Conf. Res. Comput. Mol. Biol.* **2003**, 28-37.
- [176] Berg, O.G., von Hippel, P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **1987**, *193*(4): 723-50.
- [177] Man, T.K., Stormo, G.D. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.* **2001**, *29*: 2471-2478.
- [178] Bulyk, M.L., Johnson, P.L., Church, G.M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* **2002**, *30*: 1255-1261.
- [179] Djordjevic, M., Sengupta, A.M., Shraiman, B.I. A biophysical approach to transcription factor binding site discovery. *Genome Res.* **2003**, *13*: 2381-2390.
- [180] King, O.D., Roth, F.P. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.* **2003**, *31*: e116.
- [181] Gershenson, N.I., Stormo, G.D., Ioshikhes, I.P. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Res.* **2005**, *33*: 2290-2301.
- [182] Stormo, G., Schneider, T., Gold, L., Ehrenfeucht, A. The 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.* **1982**, *10*(9): 2997-3011.
- [183] Stormo, G., Schneider, T., Gold, L. Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res.* **1986**, *14*(16): 6661-79.
- [184] Benos, P.V., Bulyk, M.L., Stormo, G.D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.* **2002**, *30*: 4442-4451.
- [185] O'Flanagan, R.A., Paillard, G., Lavery, R., Sengupta, A.M. Non-additivity in protein-DNA binding. *Bioinformatics* **2005**, *21*: 2254-2263.
- [186] Lim, L., Burge, C. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl. Acad. Sci. USA* **2001**, *98*(20): 11193-8.
- [187] Cawley, S. Statistical models for DNA sequencing and analysis spliceosome: motors, clocks, springs, and things. Cell, Statistical models for DNA sequencing and analysis. *PhD thesis. University of California at Berkeley, Berkeley, CA*, **2000**.
- [188] Xing, E., Jordan, M., Karp, R., Russell, S. A hierarchical bayesian markovian model for motifs in biopolymer sequences. In *Advances in Neural Information Processing Systems. Volume 16*. Edited by: Becker S, Thrun S, Obermayer K. MIT Press, Cambridge, MA, **2002**.
- [189] Wasserman, W., Fickett, J. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **1998**, *278*: 167-81.
- [190] Sinha, S., van Nimwegen, E., Siggia, E.D. A probabilistic method to detect regulatory modules. *Bioinformatics* **2003**, *19*(Suppl 1): i292-301.
- [191] GimaThakurta, D., Stormo, G.D. Identifying target sites for cooperatively binding factors. *Bioinformatics* **2001**, *17*(7): 608-21.
- [192] Rebeiz, M., Reeves, N., Posakony, J. SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc. Natl. Acad. Sci. USA* **2002**, *99*(15): 9888-93.
- [193] Aerts, S., Van Loo, P., Thijs, G., Moreau, Y., De Moor, B. Computational detection of cis-regulatory modules. *Bioinformatics* **2003**, *19*(Suppl 2): II5-II14.
- [194] Xing, E.P., Wu, W., Jordan, M.I., Karp, R.M. Logos: a modular bayesian model for de novo motif detection. *J. Bioinform. Comput. Biol.* **2004**, *2*: 127-54.
- [195] Gupta, M., Liu, J.S. De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* **2005**, *102*(20): 7079-84.
- [196] Frech, K., Werner, T. Specific modelling of regulatory units in DNA sequences. *Pac. Symp. Biocomput.* **1997**, 151-62.
- [197] Sharan, R., Ovcharenko, I., Ben-Hur, A., Karp, R.M. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics* **2003**, *19*(Suppl.1): i283-i291.
- [198] Scherf, M., Klingenhoff, A., Werner, T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* **2000**, *297*(3): 599-606.
- [199] Brazma, A., Vilo, J., Ukkonen, E., Valtanen, K. Data mining for regulatory elements in yeast genome. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1997**, *5*: 65-74.
- [200] Mahony, S., Hendrix, D., Golden, A., Smith, T.J., Rokhsar, D.S. Transcription factor binding site identification using the self-organizing map. *Bioinformatics* **2005**, *21*(9): 1807-1814.
- [201] Bussemaker, H.J., Li, H., Siggia, E.D. Regulatory element detection using correlation with expression. *Nat. Genet.* **2001**, *27*(2): 167-71.
- [202] Wang, T., Stormo, G.D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **2003**, *19*(18): 2369-80.
- [203] Bussemaker, H.J., Li, H., Siggia, E.D. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA* **2000**, *97*(18): 10096-100.
- [204] Gupta, M., Liu, J.S. Discovery of Conserved Sequence Patterns Using a Stochastic Dictionary Model. *J. Am. Stat. Asso.* **2003**, *98*: 55-66.
- [205] Takusagawa, K.T., Gifford, D.K. Negative information for motif discovery. *Pac. Symp. Biocomput.* **2004**, 360-71.
- [206] Sinha, S., Tompa, M. A statistical method for finding transcription factor binding sites. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2000**, *8*: 344-54.
- [207] Tompa M: An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In *Proc. Int. Conf. Intell. Syst. Mol. Biol. Heidelberg, Germany* **1999**: 262-71.
- [208] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **1993**, *262*(5131): 208-14.

- [209] Sharan, R., Ovcharenko, I., Ben-Hur, A., Karp, RM: CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics* **2003**, 19(Suppl 1): i283-91.
- [210] Curran, M.D., Liu, H., Long, F., Ge, N. Statistical methods for joint data mining of gene expression and DNA sequence database. *SIGKDD Explor Newsl* **2003**, 5(2): 122-129.
- [211] Segal, E., Yelensky, R., Koller, D. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* **2003**, 19(Suppl 1): i273-82.
- [212] Hong, P., Liu, X., Zhou, Q., Lu, X., Liu, J.S., Wong, W.H. A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics* **2005**, 21(11): 2636-2643.
- [213] Holmes, I., Bruno, W.J. Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2000**, 8: 202-10.
- [214] Bussemaker, H.J., Li, H., Siggia, E.D. Regulatory element detection using correlation with expression. *Nat. Genet.* **2001**, 27(2): 167-71.
- [215] Grad, Y.H., Roth, F.P., Halfon, M.S., Church, G.M. Prediction of similarly-acting cis-regulatory modules by subsequence profiling and comparative genomics in *D. melanogaster* and *D. pseudoobscura*. *Bioinformatics* **2004**, 20(16): 2738-2750.
- [216] Bailey, T.L., Elkan, C. The value of prior knowledge in discovering motifs with. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1995**, 3: 21-9.
- [217] Lawrence, C.E., Reilly, A.A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **1990**, 7: 41-51.
- [218] Sinha, S., Blanchette, M., Tompa, M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **2004**, 5: 170.
- [219] Prakash, A., Blanchette, M., Sinha, S., Tompa, M. Motif discovery in heterogeneous sequence data. *Pac. Symp. Biocomput.* **2004**, 348-59.
- [220] Ao, W., Gaudet, J., Kent, W.J., Muttum, S., Mango, S.E. Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* **2004**, 305(5691): 1743-6.
- [221] Liu, X., Brutlag, D.L., Liu, J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* **2001**, 6: 127-138.
- [222] Jensen, S.T., Liu, X.S., Liu, J.S., Zhou, Q. Computational Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective. *Statist. Sci.* **2004**, 19: 188-204.
- [223] Thompson, W., Rouchka, E.C., Lawrence, C.E. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.* **2003**, 31(13): 3580-5.
- [224] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, JS, Neuwald, A.F., Wootton, J.C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **1993**, 262(5131): 208-14.
- [225] Neuwald AF, Liu JS, Lawrence CE: Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* **1995**, 4(8): 1618-32.
- [226] Gupta M, Liu JS: Discovery of Conserved Sequence Patterns Using a Stochastic Dictionary Model. *J. Am. Stat. Asso.* **2003**, 94(461): 55-66.
- [227] Zhou, Q., Wong, W. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. USA* **2004**, 101(33): 12114-9.
- [228] Aerts, S., Van Loo, P., Moreau, Y., De Moor, B. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics* **2004**, 20(12): 1974-6.
- [229] Zhao, X., Huang, H., Speed, TP: Finding short DNA motifs using permuted markov models. In. In RECOMB '04- Proceedings of the eighth annual international conference on Computational molecular biology. New York, NY, USA, ACM Press, **2004**, 68-75 .
- [230] Steffen, N.R., Murphy, S.D., Toller, L., Hatfield, G.W., Lathrop, R.H. DNA sequence and structure: direct and indirect recognition in protein-DNA binding. *Bioinformatics* **2002**, 18: S22-S30.
- [231] Liu, R., Blackwell, T.W., States, D.J. Conformational model for binding site recognition by the *E.coli* MetJ transcription factor. *Bioinformatics* **2001**, 17: 622-633.
- [232] Morozov, A.V., Havranek, J.J., Baker, D., Siggia, E.D. Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.* **2005**, 33: 5781-5798.
- [233] Kaplan T., Friedman N., Margalit H. *Ab initio* prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.* **2005**, 1: e1.
- [234] Mandel-Gutfreund, Y., Margalit, H. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Res.* **1998**, 26: 2306-2312.
- [235] Kono, H., Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins* **1999**, 35: 114-131.
- [236] Havranek, J.J., Duarte, C.M., Baker, D. A simple physical model for the prediction and design of protein-DNA interactions. *J. Mol. Biol.* **2004**, 344: 59-70.
- [237] Benham, C.J. Computation of DNA structural variability—a new predictor of DNA regulatory regions. *Comput. Appl. Biosci.* **1996**, 12: 375-381.
- [238] Quandt, K., Grote, K., Werner, T. GenomeInspector: a new approach to detect correlation patterns of elements on genomic sequences. *Comput. Appl. Biosci.* **2003**, 12: 405-413.
- [239] Pedersen, A.G., Baldi, P., Chauvin, Y., Brunak, S. The biology of eukaryotic promoter prediction—a review. *Comput. Chem.* **1999**, 23(3-4): 191-207.
- [240] Ponomarenko, J., Ponomarenko, M., Frolov, A., Vorobyev, D., Overton, G., Kolchanov, N.A. Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics* **1999**, 15(7-8): 654-68.
- [241] El Hassan, M.A., Calladine, C. Conformational characteristics of DNA: empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *Roy. Soc. of London Phil. Tr. A* **1997**, 355(1722): 43-100.
- [242] Burden, S., Lin, X. and R. Zhang, R. Improving promoter prediction Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences. *Bioinformatics* **2005**, 21(5): 601-607.
- [243] Liu, R. and States, D. Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. *Genome Res.* **2001**, 12: 462-469.
- [244] Mellor, J. The dynamics of chromatin remodeling at promoters. *Mol. Cell.* **2005**, 19: 147-157.
- [245] Pudimat, R., Schukat-Talamazzini, E., Backofen, R. Feature Based Representation and Detection of Transcription Factor Binding Sites. *Proc. German Conf. Bioinform.* **2004**, 43-52.
- [246] Perkins, D., Jeffries, C., Sullivan, P., Expanding the 'central dogma': the regulatory role of nonprotein coding genes and implications for the genetic liability to schizophrenia. *Mol. Psychiatry* **2005**, 10(1): 69-78.
- [247] Szymanski, M., Barciszewski, J. RNA regulation in mammals. *Ann. N. Y. Acad. Sci.* **2006**, 1067: 461-8.
- [248] Qi, L., Li, X., Zhang, S., An, D. Genetic regulation by non-coding RNAs. *Sci. China C Life Sci.* **2006**, 49: 201-17.
- [249] Yang, L., Ke, Y. Noncoding RNA, a new focus of functional genomic study. *Beijing Da Xue Xue Bao* **2006**, 18(38): 444-6.
- [250] Mattick, J., Makunin, I. Non-coding RNA. *Hum. Mol. Genet.* **2006**, 15: R17-R29.