Computational analysis of promoters and DNA-protein interactions

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Andrija Tomovic

aus Belgrad, Serbien

Basel, 2009

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Andreas Engel
Prof. Dr. Torsten Schwede
Prof. Dr. Patrick Matthias
Dr. Edward J. Oakeley

Basel, 19.2.2008

Prof. Dr. Hans-Peter Hauri
Dekan

ABSTRACT

The investigation of promoter activity and DNA-protein interactions is very important for understanding many crucial cellular processes, including transcription, recombination and replication. Promoter activity and DNA-protein interactions can be studied in the lab (*in vitro* or *in vivo*) or using computational methods (*in silico*). Computational approaches for analysing promoters and DNA-protein interactions have become more powerful as more and more complete genome sequences, 3D structural data, and high-throughput data (such as ChIP-chip and expression data) have become available. Modern scientific research into promoters and DNA-protein interactions represents a high level of co-operation between computational and laboratorial methods.

This thesis covers several aspects of the computational analysis of promoters and DNA-protein interactions: analysis of transcription factor binding sites (investigating position dependencies in transcription factor binding sties); computational prediction of transcription factor binding sites (a new scanning method for the *in silico* prediction of transcription factor binding sites is described); computational analysis of crystal structures of DNA-protein interactions (multiple proteins bound to DNA); and computational predictions of transcription factor co-operations (investigating dependencies between transcription factors in human, mouse and rat genomes, and a new method of *in silico* prediction of cis-regulatory motifs and transcription start sites is described). In addition, this thesis reports how one statistical method for the analysis of transcription factor binding sites can be used for estimating the quality of multiple sequence alignments.

The main finding reported in this thesis is that it is wrong to assume, *a priori*, that positions in transcription factor binding sites are all either independent or dependent on one another. Position dependencies should be tested using rigorous statistical methods on a case-by-case basis. When dependencies are detected, they can be modelled in a very simple way, which doesn't require complex mathematical tools with a lot of parameters and more data. An example of such a model, including a web-based implementation of the algorithm, is reported in this thesis. It has also been shown that the conformational

energy (indirect readout) of DNA in complexes with transcription factors which have dependent positions in their binding sites is significantly higher than in those with transcription factors which do not have dependent positions in their binding sites.

The structural analysis of multiple protein-DNA interactions showed that the formation of interactions between multiple proteins and DNA results in a decrease in protein-protein affinity and an increase in protein-DNA affinity, with a net gain in overall stability of complexes where multiple proteins are bound to DNA. This effect is clearly important for modelling transcription factor co-operativity. In addition, the physical overlap of two factors does not simply relate to the region on the DNA where the binding site is found. Two factors may lie very close together but possibly not physically overlap because their side-chains can interlink with one another. In this way, it is possible to find a large overlap between two transcription factor binding sites, but from a 3D perspective it is still possible for both factors to bind simultaneously. It may also be that one transcription factor binds to the minor and another to the major groove of DNA. That information is also useful for modelling transcription factor co-operativity.

Moreover, this thesis reports the results from a computational prediction of dependencies (co-operativities) between transcription factors which usually act together in gene regulation in human, mouse and rat genomes. It is shown that that the computational analysis of transcription factor site dependencies is a valuable complement to experimental approaches for discovering transcription regulatory interactions and networks. Scanning promoter sequences with dependent groups of transcription factor binding sites improve the quality of transcription factor predictions. Finally, it has been demonstrated that modelling transcription factor co-operativities improves the quality of transcription start site predictions. For three genes (ctmp, gap-43 and ngfrap) *in-vivo* validation of the predicted transcription start sites is performed.

Finally, the Bayesian method for the detection of dependencies between positions in transcription factor binding sites can easily be converted into a method for estimating the quality of multiple sequence alignments. That method is simple, linear complexity, which is easy to implement and which performs better than other state-of-the-art methods which are more complex.

ORIGINAL PUBLICATIONS

I

Andrija Tomovic and Edward J. Oakeley. Position dependencies in transcription factor binding sites. *Bioinformatics 2007, 23(8):933-941*.

II

Andrija Tomovic and Edward J. Oakeley. Quality estimation of multiple sequence alignment by Bayesian hypothesis testing. *Bioinformatics 2007, 23(18):2488-2490.*

III

Andrija Tomovic and Edward J. Oakeley. Computational structural analysis: multiple proteins bound to DNA. *PLoS ONE. 2008 Sep 19;3(9):e3243.*

IV

Andrija Tomovic, Michael Stadler and Edward J. Oakeley. Transcription factor site dependencies in human, mouse and rat genome. *BMC Bioinformatics 2009, 10:33.*

CONTENTS

LIST OF ABBREVIATIONS

bp	base pair
DBD	DNA-binding domain
DNase	deoxyribonuclease
DPE	downstream promoter element
CAGE	cap analysis of gene expression
cDNA	complementary deoxyribonucleic acid
ChIP	chromatin immunoprecipitation
ChIP-chip	chromatin immunoprecipitation microarrays
CRM	cis-regulatory module
DNA	deoxyribonucleic acid
EMSA	electrophoretic mobility shift assay
FN	false negative
FP	false positive
HMM	hidden Markov model
kbp	kilo base pairs
mRNA	messenger ribonucleic acid
NMR	nuclear magnetic resonance
PDB	protein data bank
PET	paired-end ditag technology
PWM	position weight matrix
RACE	rapid amplification of cDNA ends
RNA	ribonucleic acid
RNase	ribonuclease
SSD	signal-sensing domain
TAD	trans-activating domain
TF	transcription factor
TFBS	transcription factor binding site

TN	true negative
TP	true positive
TSS	transcription start site
UV	ultraviolet
UCS	upstream control sequence

1. Introduction

Computational techniques in molecular biology can be useful from both a theoretical and a practical point of view. From the theoretical point of view, computational methods can help to mine the huge amounts of data produced in the laboratory, in order to characterise the data, and find interesting patterns, clusters and rules. With the current expansions in biotechnology, the amount of high-throughput and other laboratory data increases every day, and the need for the mining of these data is increased. From the practical point of view, computational methods may be useful for different kinds of predictions and simulations, or for assistance in the laboratory. This is crucial for saving time, money and resources in laboratory research. Modern scientific research into promoters and DNA-protein interactions represents a high level of co-operation between computational and laboratorial methods.

1.1 DNA-binding proteins

DNA-binding proteins are important for the regulation of many crucial cellular processes (including gene expression, recombination, translation and replication). Because of that, it is very important to investigate DNA-binding proteins and understand the DNA-binding process. There are several kinds of DNA-binding proteins:

- **Transcription factors** are regulatory DNA-binding proteins which play a crucial role in the regulation of gene expression. The total number of transcription factors in an organism increases with the number of genes in the genome [1] and with the size of the genome (there is a power-law relationship between genome size and total number of transcription factors as $N \sim G^{1.9}$ for prokaryotes and $N \sim G^{1.3}$ for eukaryotes, where N is the total number of transcription factors and G the number of genes) [2, 3]. The genome sequences of *C. elegans* and *Drosophila* reveal at least 1,000 transcription factors [4, 5]. There are probably 3,000 transcription factors in humans [6]. Yeast contains an average of one transcription factor per 20

genes, while humans appear to contain one factor for every ten genes [1]. Transcription factors can be activators of transcription processes, but they can also act by inhibiting the transcription of specific genes [7]. Based on this, we can separate transcription factors into two classes: *transcription factor activators*; and *transcription factor inhibitors (repressors)*. Transcription factors bind to short DNA sequences known as transcription factor binding sites (TFBS, DNA-binding motifs, cis-regulatory elements).Transcription factor binding sites are usually very short and highly degenerate. It is possible to distinguish basal transcription factors and enhancer transcription factors based on the position of their DNA-binding motifs on the promoter. The part (domain) of the transcription factor that binds to DNA is called the transcription factor DNA-binding domain. Transcription factors can be classified according to the structural similarity of their DNA-binding domains (DBD) [8][1]. Some well characterised DNA-binding domains include: the helix-turn-helix motif (found in homeobox transcription factors); the two cysteine-two histidine zinc finger (found in the Sp transcription factor family); the multi-cystine zinc finger (found in the steroid-thyroid hormone receptor family); and the Ets domain [7]. Apart from the DNA-binding domains, transcription factors usually contain a trans-activating domain (TAD) which contains binding sites for other proteins (transcription co-regulators) [7, 9]. In addition, transcription factors sometimes have a signal-sensing domain (SSD) (e.g. a ligand-binding domain) which senses external signals and, in response, transmits these signals to the rest of the transcription complex, resulting in up- or down-regulation of gene expression [7]. Very often, the TAD and SSD are the same. In order to act as transcription activators or repressors, very often transcription factors should be activated (or deactivated) through their SSD by ligand binding (like nuclear receptors), interactions with other transcription factors (making cis-regulatory modules), the binding of co-regulators and phosphorylation [10]. Transcription factor activators can also be classified based on their function [11]:

[1] In the following text, the classification of all DNA-binding proteins is going to be based on the structural analysis of their DNA-binding motifs.

I. *constitutively active* - present in all cells at all times - general transcription factors, Sp1, CCAAT-binding protein, NF1 and many others;

II. *regulatory transcription factors*

II.A *developmental (cell-specific)* - expression is tightly controlled, but they require no additional activation once expressed - GATA, HNF, PIT-1, MyoD, Myf5, Hox, winged helix;

II.B *signal-dependent* - requires external intra- or extracellular signal for activation

II.B.1 *the steroid receptor superfamily* (extracellular ligand dependent - nuclear receptors);

II.B.2 *transcription factors activated by internal (cell-autonomous) signals* (intracellular ligand-dependent - activated by small intracellular molecules - SREBP, p53, orphan nuclear receptors);

II.B.3 *transcription factors activated by cell-surface receptor-ligand interactions* (cell membrane receptor-dependent - second messenger signalling cascades resulting in the phosphorylation of the transcription factor);

II.B.3.a *constitutive nuclear factors activated by serine phosphorylation* (reside in the nucleus regardless of activation state, e.g. CREB, AP-1, Mef2);

II.B.3.b *latent cytoplasmic factors* (inactive forms reside in the cytoplasm but when activated are translocated into the nucleus, e.g. STAT, R-SMAD, NF-kB, Notch, TUBBY, NFAT);

The repression of gene expression can occur by the transcription factor repressor binding to DNA and preventing an activator from binding and activating the transcription process, by the transcription factor repressor interacting with the activator and in that way preventing its DNA from binding, by the repressor binding to DNA with the activator and neutralising its ability to activate transcription, or by direct repression by inhibiting the transcription factor [7]. Because transcription factors play an important role, it is not unexpected that

alterations in them can result in human diseases [7, 12]. Such diseases can be divided into three major groups: developmental disorders, disorders of hormone responses and cancer [7].

- **Histones** are DNA-binding proteins responsible for the first, and most basic, level of chromosome organisation, the nucleosome, which was discovered in 1974 [13]. Histones are present in huge quantities in the cell (about 60 million molecules of each type per human cell) [13].The structural organisation of nucleosomes was determined after isolating them from unfolded chromatins using nucleases [13]. The nucleosome core particle consists of an octomer complex of eight histone proteins (two molecules of each of histones H2A, H2B, H3 and H4) and double-stranded DNA (~146 bp long) wrapped around the octomer (Figure 1). Each nucleosome core particle is separated from the next by a region of linker DNA (which can vary in length from 0 up to about 80 bp, depending on the species [14]). The term nucleosome refers to a nucleosome core particle plus one of its adjacent DNA linkers. Nucleosomes are the first level of DNA packing (compressing DNA to about one-third of its initial length). There are indications that nucleosome organisation is encoded in eukaryotic genomes, i.e. that genomes use nucleosome sequence preference to control the distribution of nucleosomes *in vivo* in a way that strongly impacts on the ability of non-histone DNA binding proteins to access particular binding sites [15]. According to this statement, remodelling factors do not themselves determine the destinations of the nucleosomes that they mobilise. An array of nucleosomes, together with histone H1 molecules, is known as “beads on a string” and represents the second level of chromosome organisation. Histone H1 is larger than the core histones and is considerably less well conserved. Further nucleosome arrays are usually packed together into quasi-regular arrays to form a 30-nm fibre (solenoid, chromatin fibre). The next level of chromosome organisation is euchromatin and heterochromatin. Euchromatin makes up most of the interphase chromosomes, and probably corresponds to looped domains of 30-nm fibres. Euchromatin is

interrupted by heterochromatin, on which 30-nm fibres are subjected to additional levels of packing, and this usually renders it resistant to gene expression [13].

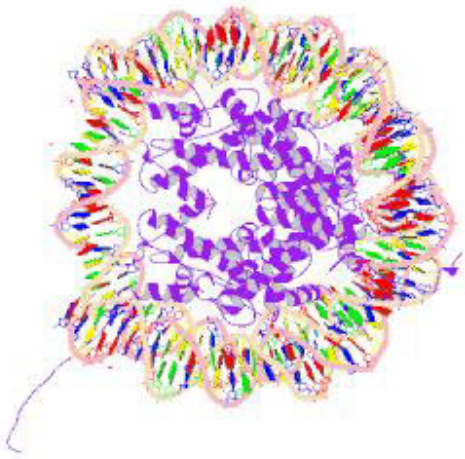


Figure 1. *Complex between a nucleosome core particle (octomer) and a 146bp DNA fragment (source: 1aoi.pdb)*

- **DNA-modifying enzymes** such as:
    - **Nucleases**, which are enzymes that cleave the phosphodiester bonds between the nucleotide subunits of nucleic acids (i.e. catalyse the hydrolysis of nucleic acids [16]). Earlier, they were marked with the term "polynucleotidase" or "nucleodepolymerase" [17]. Nucleases have an important biological role but, in addition to that, they are used in the laboratory for recombinant DNA technology, molecular cloning and genomics. Nucleases are further described as endonucleases or exonucleases. Endonucleases break nucleic acid chains somewhere in the middle of a molecule, rather than at the ends. Exonucleases remove nucleotides from the ends of the molecule. There are many types of nucleases that have been isolated and characterised. Some of the more widely used nucleases are [18]:
        i. *deoxyribonuclease I (DNase I)* - an endonuclease that cleaves double-stranded or single-stranded DNA (does not cleave RNA). This is the most widely used nuclease and is purified from bovine

pancreas. Cleavage preferentially occurs adjacent to pyrimidine (C or T) residues, and major products are di-, tri- and tetranucleotides. Common applications of DNase I are: eliminating DNA (e.g. plasmid) from preparations of RNA; analysing DNA-protein interactions via DNase footprinting; and nicking DNA prior to radiolabelling by nick translation.

ii. *exonuclease III* - the nuclease that removes mononucleotides from the 3' termini of duplex DNA. This nuclease is purified from *E. coli* and frequently used to prepare a set of nested deletions of the termini of linear DNA fragments.

iii. *mung bean nuclease* - a nuclease that digests single-stranded DNA to 5'-phosphorylated mono- or oligonucleotides. This nuclease is purified from mung bean sprouts and frequently used to remove single-stranded 5' extensions from DNA (or RNA), leaving blunt, ligatable ends.

iv. *nuclease S1* - a nuclease that, in low concentrations, digests single-stranded DNA or RNA, while in high concentrations digests double-stranded nucleic acids (DNA:DNA, DNA:RNA or RNA:RNA). This nuclease is purified from *Aspergillus* and frequently used to analyse the structure of DNA:RNA hybrids (S1 nuclease mapping), and to remove single-stranded extensions from DNA to produce blunt ends.

- **Polymerases** are enzymes which synthesise polynucleotide chains from nucleoside triphosphates. They function by adding nucleotides onto the 3' hydroxyl group of the previous nucleotide in the DNA strand and work from the 5' to the 3' end [19].
- **DNA integrases** are enzymes produced by a retrovirus that helps in the integration of its genetic material into the DNA of infected cell [20].
- **Helicases** are enzymes which use the chemical energy in nucleoside triphosphates to break hydrogen bonds between bases and unwind the DNA double-helix into single strands [21].

- **Topoisomerases, ligases, DNA methylases** and others.

Thanks to an increased number of available 3D structures, it is now possible to analyse DNA-protein interactions from the structural point of view. In this way, a lot of valuable information about the general features of such complexes has been discovered [22-28].
In addition, DNA-binding proteins have been classified based on the structures of the DNA-binding regions in the proteins [24]. There are several main structural classes of DNA-binding proteins:

i. Helix-turn-helix proteins. This group of proteins has a characteristic DNA-binding motif which contains 20 amino acids of two almost perpendicular α helices connected by a four-residue β turn (Figure 2) [29]. Many prokaryotic and eukaryotic transcription factors and enzymes belong to this class [30]. Helix-turn-helix proteins bind to the major groove of DNA [29]. The prokaryotic transcription factors from this class bind to palindromic DNA sequences such as homodimers. Eukaryotic proteins from this class, such as members of the homeodomain family, bind both as monomers and heterodimers to non-symmetrical target sites. There are 16 homologous families in this class [24].

Figure 2. *Crystal structure of the lambda repressor-operator complex (source: 1lmb.pdb), as an example of a helix-turn-helix DNA-binding protein*

ii. Zinc-coordinate proteins. Proteins in this class have a DNA-binding motif which is characterised by the tetrahedral co-ordination of one or two zinc ions by

conserved cysteine and histidine residues (Figure 3) [30]. This is the largest single class of eukaryotic transcription factors [29]. There are four homologous families in this group (the ββα zinc-finger family, the hormone receptor family, the loop-sheet-helix family and the gal4 family) [29].

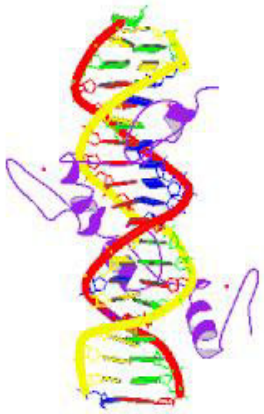


Figure 3. *Crystal structure of the human YY1 zinc finger (source: 1ubd.pdb), as an example of a zinc-coordinate DNA-binding protein*

iii. Zipper-type proteins. This class of DNA-binding proteins derives its name from the method of dimerisation used by its members (Figure 4) [29]. This class contains only eukaryotic DNA-binding proteins in two homologous families (leucine zipper family and helix-loop-helix proteins) [29]. The DNA binding site is pseudo-symmetrical, and typically eight base-pairs long.

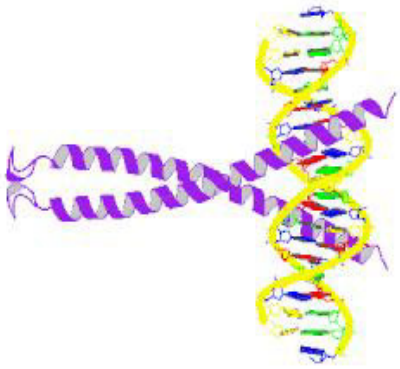


Figure 4. *Crystal structure of GCN4-BZIP (source: 1dgc.pdb), as an example of a zipper-type DNA-binding protein*

iv. Other α-helix proteins. This class contains seven homologous families and eukaryotic and prokaryotic DNA-binding proteins. All proteins from this class use α helices as the main method of DNA binding (Figure 5) [29].

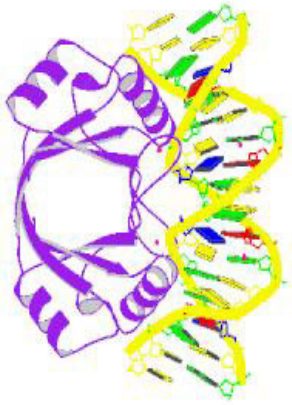


Figure 5. *Crystal structure of the bovine papillomavirus-1 E2 DNA-binding domain (source: 2bop.pdb), as an example of another α-helix DNA-binding protein*

v. β-sheet proteins. DNA-binding proteins from this class use β-strand structures for DNA recognition and binding (Figure 6). This class only contains the TATA box-binding protein family, which is characterised by the use of a wide β-sheet to bind the DNA. A ten-stranded anti-parallel β-sheet, which joins the domains, covers the DNA minor groove[29].

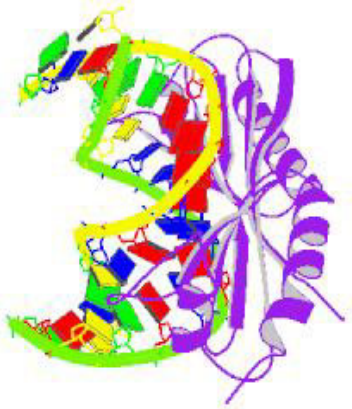


Figure 6. *Crystal structure of the human TBP core domain (source: 1cdw.pdb), as an example of a β-sheet DNA-binding protein*

vi. β-hairpin/ribbon proteins. DNA-binding proteins from this class are different from the TATA box-binding proteins in that they use smaller, two- or three-stranded β-sheets or hairpin motifs to bind in either the DNA major or minor grooves [29] (Figure 7). This class contains six homologous families and eukaryotic and prokaryotic DNA-binding proteins.

vii. Other DNA-binding proteins. This class contains two non-enzymatic homologous families which do not use any well defined secondary structural motifs for DNA binding. This class contains only eukaryotic DNA-binding proteins.

viii. Enzymes. This class is separated from the other classes because it contains DNA-binding proteins that have no common structural motifs for binding DNA, but which are brought together on the basis of their functions (all alter DNA structure through the catalysis of a chemical process) [29]. This class contains eukaryotic and prokaryotic DNA-binding proteins, and these proteins use an extensive combination of α-helices, β-strands and loops to recognise and bind DNA [29].

Figure 7. *Crystal structure of the met repressor-operator (source: 1cma.pdb), as an example of a β-hairpin/ribbon protein DNA-binding protein*

Identification and analysis of DNA-binding proteins (and transcription factors) and their binding sites can be performed in the laboratory (*in vivo* or *in vitro*) and using computational techniques (*in silico*).

1.2 Promoters

Promoters can be defined as the genomic regions that surround a transcription start site (TSS)[2] or cluster of TSSs [31]. There is no precise definition of promoter length. Usually, it is defined empirically as the DNA region which is required to recruit the transcription initiation complexes and initiate transcription, together with external signals such as enhancer transcription factors [31].

It is possible to distinguish a core (or basal) promoter from an enhancer promoter (upstream promoter region). A core or basal promoter is a DNA region where basal transcription factors (basal machinery) bind. The enhancer promoter is a DNA region where additional transcription activators bind. Enhancers were first identified in viruses and then in cellular genes. Transcriptional repressors (transcriptional silencers), which repress the transcription process, can also bind in that region.

There is a difference in transcription (and promoter) complexity between bacteria and eukaryotic organisms [13]. In the bacterial nucleus, there is only one type of RNA polymerase, and the key motif in promoters is the pribnow box. In the presence of the σ-factors, bacterial RNA polymerases can recognise bacterial promoters without the help of any other transcription factors [32]. In contrast, eukaryotic nuclei have three RNA polymerases:

i. RNA polymerase I (Pol I)
ii. RNA polymerase II (Pol II)
iii. RNA polymerase III (Pol III)

[2] Transcription start site (TSS) is a nucleotide in the genome that is the first to be transcribed into a particular RNA [31].

RNA polymerases I and III transcribe the genes that encode transfer RNA, ribosomal RNA and various small RNAs. RNA polymerase II transcribes all other genes, including all those that encode proteins [13]. Because of these RNA polymerases, there are three different classes of promoters in eukaryotic nuclei:

i. Pol I promoters
ii. Pol II promoters
iii. Pol III promoters

Pol I interacts with Pol I promoters complexed with UCS and a second factor (variously named SL1, TIFIB, D or Rib1). An Upstream binding protein (UBF) binds to the UCS and recruits TATA binding protein (TBP) together with the TBP associated factors (TAFs). Rm3/TIF-IA get phos and binds to Pol I then Pol I binds to UBF/SL1 via Rm3/TIF-IA. Pol II binds to Pol II promoters with basal transcription factors (TFIID, A, B, E, F, H and J) and different upstream (enhancer) transcription activators for different individual promoters (Figure 8). Pol II promoters contain TATAAA consensus sequence, called a *TATA box*or a *Hogness box* (the spacing between the TATA box and the initiator is 25bp in all eukaryotes except plants where it is 35 bp). Pol III binds to most Pol III promoters with TFIIIB and C, but to the 5S gene promoters with TFIIIA as well [33].

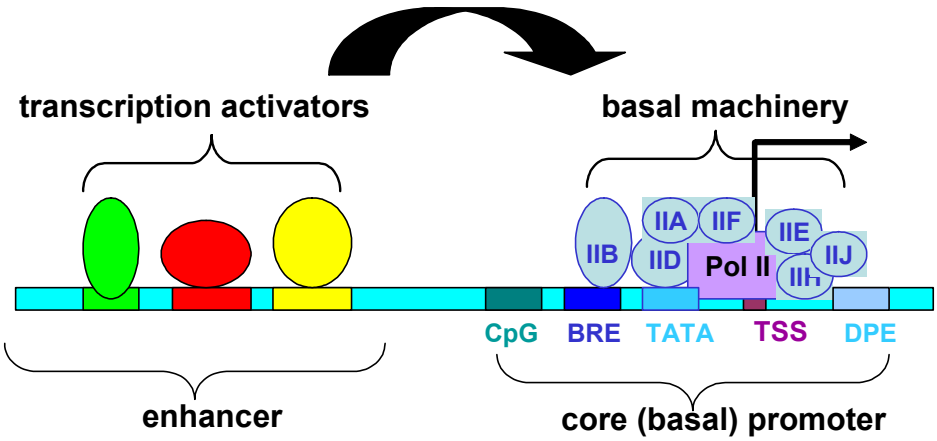


Figure 8. *Pol II promoter organisation: the core promoter often contains a CpG island, a TATA box (TFIID-binding element), a BRE (TFIIB-binding element), a DPE (downstream promoter element) and an initiator element (at the TSS). The enhancer contains transcription activator binding sites (modules)*

There is an experimentally confirmed situation where genes lie on opposite strands, with their TSSs in close proximity with each other, to form so-called bidirectional promoters [34]. Trinklein et al. estimated that 1,352 gene pairs in the human genome have TSSs on the opposite strand that are less than 1 kbp away. In the mouse genome it is estimated that there are 1,638 gene pairs that have TSSs on the opposite strand that are separated by less than 1 kbp [35].

Very often in mouse and human genomes, protein-coding genes are associated with more than one promoter region [31, 36]. Most well-supported alternative promoters are found at the 5' ends of known cDNAs, or in protein-coding exons [31].

Analysis of promoters includes promoter identification, enhancer-promoter communications, TSS identification, and analysis of DNA methylation. Promoter analysis can be performed in the laboratory (*in vivo* or *in vitro*) and using computational techniques (*in silico*).

1.3 Laboratory techniques for promoter and DNA-protein interaction analysis

There are two groups of laboratory methods for TSS identification:

- Methods based on sequencing of cDNA
    i. RACE - rapid amplification of cDNA ends [37]. This method is used to detect the 5' ends of individual RNAs and is useful for targeting particular loci of interest with higher scalability than hybridisation-based methods. However, this is a low-throughput method (different primers must be used for different methods), and information about the span of the full transcript is not retained.
    ii. 5' tag sequencing [38], exemplified by the cap analysis of gene expression (CAGE) technique. These methods have the highest throughputs, but information about the span of the full transcript is not retained.

iii. 5'-3' tag sequencing [31], exemplified by paired-end ditag technology (PET). Because both 5' and 3' ends are sequenced, more information is available and this method can be used together with ChIP to sequence DNA that is bound by a factor of interest. However, this method is a lower-throughput method than 5' tagging.

iv. next generation sequencing systems (ultrahigh-throughput methods) like Solexa, ABI and 454 [39].

- Methods that involve hybridisation of RNA or cDNA to DNA probes [31]:
  i. Nuclease protection methods [40] rely on hybridising a labelled DNA probe. These methods are designed to be complementary to a postulated TSS region, with a source of mRNA, and incubating with a nuclease (often S1 nuclease) that cleaves single-stranded molecules. Methods are gel-based, low-throughput and independent of reverse transcription, and require the use of radioisotopes in order to be best done.
  ii. Primer extension methods use a labelled primer that is complementary to an internal region of an mRNA used for reverse transcription. Methods are gel-based and low throughput, and require detection with radioisotopes [31].
  iii. Tilling arrays provide a snapshot of all the transcribed regions in the genome, not only the 5' or 3' ends. Exon boundaries may be observed on high density tiling arrays (e.g. Affymetrix) with high precision (within 35bp). The exact splice point is then easy to find by looking for splice junctions within the region of interest. Alternative-splicing information can also be distinguished by this technique because it provides a large number of signal measurements for each exon. The results represent integrated signals from all transcript variants but many into a single signal that must be deconvolved. Algorithms to do this more efficiently are under active development.

There are several laboratory techniques for assessing DNA-protein interactions, including:

- Electrophoretic mobility shift assay (EMSA) [41, 42] where the binding of a sequence-specific DNA bound to a radioactively labelled DNA fragment gives the DNA-protein complex with a reduced mobility of the DNA in a non-denaturing polyacrylamide gel [43].
- DNase I protection (footprinting) assay [44, 45] where the binding of a protein to a specific region within a singly end-labelled DNA fragment protects it from digestion by DNase I [43].
- Methylation interference assay [46], which is based on the fact that methylation of specific guanine or adenine residues within the target DNA sequence inhibits the binding of a transcription factor to that site [43].
- UV cross-linking [47], which is based on the fact that when a protein-DNA complex is irradiated with UV light, it causes the formation of covalent bonds between pyrimidines and certain amino acid residues in the transcription factor that are in close proximity to the DNA [43].
- Southwestern blotting [48], which is based on the fact that cell extracts containing the DNA-binding protein are resolved by denaturing polyacrylamide gel electrophoresis followed by electrophoretic transfer to a nitrocellulose membrane [43].
- Chromatin immunoprecipitation (ChIP) assays and ChIP-chip (chromatin immunoprecipitation microarrays) methods [49, 50]. These methods are used to isolate DNA fragments that are bound to DNA-binding proteins or their complexes. They are especially useful when the protein of interest is known. ChIP assays capture *in vivo* DNA-protein interactions by cross-linking proteins to their DNA binding sites using formaldehyde. First, the DNA is fragmented into small pieces of 100-500 bp (average), and after that precipitation is done by transcription factor specific antibody. Finally, reversal of the cross-linking reaction releases the DNA for subsequent detection by PCR amplification [50]. In order to find where the protein binds across the whole genome, ChIP-chip can be used as a combination of a ChIP assay and tiling microarray (chip). ChIP-seq is a variant of this where the fragments from ChIP are sequenced in a next generation sequencer (for example Solexa or ABI). DNA that has undergone a ChIP assay

may be labelled with the fluorophore Cy5. Its signal, when bound to an array of target sequences, is compared with the signal of an equal amount of total input DNA which is labelled with Cy3 (e.g. Nimblegen arrays). Alternatively this may be done with a single sample hybridisation to an array where biotin is incorporated and detected with streptavidin-phycoerithrin (Affymetrix) In order to identify binding sites, one should compare the relative enrichment of immunoprecipitated DNA over total input DNA [50].

- Transfection assays [50]. Different type of plasmids (with different kinds of DNA binding elements) can be transfected separately into cultured cells and, after that, the activity of a reporter enzyme can be noted.
- Proximity-dependent DNA ligation assays [51].

In order to study the biochemical properties of transcription factors, it is very often necessary to study them in pure (cloned) forms. There are two major categories for the purification and cloning of transcription factors [43]:

i. *biochemical purification of transcription factors*
ii. *expression cloning of transcription factors* (*in situ* detection of transcription factors, the yeast one-hybrid selection systems)

3D crystal structure of DNA-protein interactions can be constructed by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy (for small proteins). Crystal structures of DNA-protein complexes give precise information about the positioning of the proteins relative to the double helix (DNA).

The laboratory methods for studying protein-DNA interactions and promoters are very useful, and are important for getting more knowledge about cellular processes such as transcription, recombination and replication. These methods produce a lot of information which cannot be analysed by simple observation, and therefore precise computational techniques should be applied. In this way, it is possible to understand the data produced by laboratory methods better. In addition, laboratory methods are usually very expensive

and time-consuming, and it is therefore possible to save time and resources by doing *in silico* research in combination with laboratory methods.

1.4 Computational method for promoter and DNA-binding protein analysis

There are several computational approaches for studying promoters and their DNA-binding proteins, such as:

I. computational predictions of promoters, DNA methylation sites and TSSs;
II. computational predictions of binding sites of DNA-binding proteins;
III. computational structural analysis of DNA-protein crystal complexes.

Computational approaches for analysing promoters have become more powerful as more and more complete genome sequences, ChIP data, 3D structural data and expression data have become available. The computational prediction of promoter regions and transcription start sites is still in its infancy; one of the main problems is that the promoter is defined functionally rather than structurally, which greatly limits the success of attempts to model it [52]. Some tools for the prediction of promoter regions or starts of transcription have already been published, including: McPromoter [53]; FunSiteP [54]; Dragon Promoter Finder [55]; Core-promoter [56]; WWW PromoterScan [57]; Promoter 2.0 [58]; NNPP [59]; and FirstEF [60].

The computational prediction of transcription factor binding sites is also an open-research problem. The main problem is that binding sites for transcription factors are typically short and highly degenerate. Identification of such sequences in the promoter is not easy, because such short sequences are expected to occur at random every few hundred base pairs. So, the question is how to separate real motifs from false positives [61]. Methods for the computational prediction of transcription factor binding sites can be separated in two groups:

i. scanning methods (inferring binding specificities from known binding sites, examples of tools based on these methods including: MATCH [62], ConSite [63], MAPPER [64] and rVista [65])

ii. *ab initio* methods (inferring binding specificities without a prior knowledge of binding sites, examples of tools based on these method including: Gibbs sampler [66], MEME [67], Bioprospector [68] and YMF [69]).

Computational representation of transcription factor binding sites (cis-regulatory motifs and DNA motifs) can be performed in two ways:

- assuming that each base in the binding site occurs independently (Figure 9). Models based on this premise include [70, 71]:

    1. word (search for exact sequence match);
    2. consensus sequence (pattern representation, regular expression, average sequence form multiple binding sites);
    3. matrix profile (position frequency matrix, position weight matrix);
    4. sequence logos

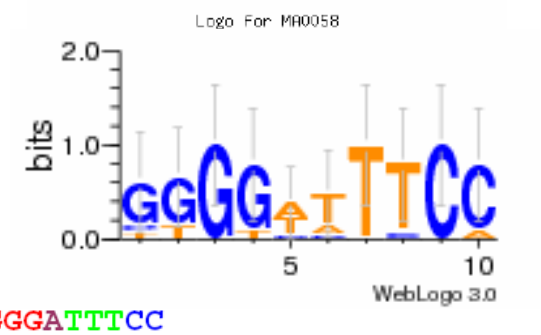
- incorporating dependencies between positions in transcription factor binding sites:

    1. Bayesian networks [72]
    2. optimised mixed Markov models [73], HMMs [64] and other Markov model variants [74]
    3. graph-based methods [75]
    4. generalised weight-matrix models and weight-array models [76]
    5. non-parametric methods [77].

Multiple alignment of 9 binding sites of transcription factor MAX (MA0058-JASPAR database):

GGGGATTTCC
GGGGATTTCC
GGGGAATTCC
GGGGTTTTCC
GGGGCATTCC
GTGGTTTTCC
TGGGACTTCC
CTGGTTTTCC
GGGTTTTCCA

Sequence logo:



Consensus sequence: GGGGATTTCC

Alternate consensus sequence regular expression: BKGKHTYCM

Position Frequency matrix:

	1	2	3	4	5	6	7	8	9	10
A	0	0	0	0	4	2	0	0	0	1
C	1	0	0	0	1	1	0	1	9	8
G	7	7	9	8	0	0	0	0	0	0
T	1	2	0	1	4	6	9	8	0	0

Position Weight matrix:

	1	2	3	4	5	6	7	8	9	10
A	-inf	-inf	-inf	-inf	0.47	-0.22	-inf	-inf	-inf	-0.91
C	-0.91	-inf	-inf	-inf	-0.91	-0.91	-inf	-0.91	1.28	1.16
G	1.02	1.02	1.28	1.16	-inf	-inf	-inf	-inf	-inf	-inf
T	-0.91	-0.22	-inf	-0.91	0.47	0.87	1.28	1.16	-inf	-inf

Figure 9. *Examples of* in silico *representation of transcription factor binding sites, assuming that each base in the binding site occurs independently*

There are several strategies for improving the accuracy of *in silico* methods (both scanning and *ab initio* methods) for transcription factor binding site predictions:

- Using structural information of transcription factors [78-80]: it is known that similar transcription factors bind in a similar way to DNA. Some DNA-binding proteins from the same family recognise binding sites which have similar length, symmetry and specificity [80].

- Using comparative genomic-phylogenetic footprinting [81, 82]: sequence similarities resulting from selective pressure during evolution is a basic principle for many bioinformatical methods [83]. Key assumptions in the application of phylogenetic footprinting are: that the regulation of orthologous genes is controlled in the same way in different species; and that mutations within functional regions of genes will accumulate more slowly than mutations in regions with no sequence-specific function [70].
- Using information about nucleosome occupancy [84]: Segal et al. [15] reported, recently, a nucleosome-DNA interaction computational model which can be used to predict transcription factor binding sites taking into consideration that positions which are occupied by nucleosomes are not accessible for transcription factors.
- Using models which assume dependencies between positions in transcription factor binding sites: it has been reported that methods which incorporate dependencies between positions in transcription factor binding sites predict binding sites more accurately (lower false positive rates) but, on the other hand, require a more complex mathematical approach (more parameters to estimate) and more data. However, a method is shown in this thesis for modelling position dependencies in a simple way that does not require complex mathematical models or any more data than models which assume independence of positions in transcription factor binding sites [85].
- Using modelling of co-operativity between transcription factors (combinatorial interactions between transcription factors) [86-88]: it is very well known that transcription factors (specially in eukaryotes) rarely act alone in regulating gene expression. In most cases, multiple factors bind DNA, sometimes in close proximity with each other, forming cis-regulatory modules (CRMs) [71].

Possible future work for improving *in silico* predictions could be in the field of DNA methylation and/or using functional information from transcription factors (perhaps transcription factors with the same function (section 1.1) bind to DNA in similar ways). And, finally, further work could include the construction of a unified framework which will unite all the previously mentioned strategies.

Computational structural analysis of DNA-protein crystal structures is actually data-mining on a dataset of 3D structures. Thanks to the increased number of available 3D structures of DNA-protein interactions stored in the PDB (Protein Data Bank) database [89], this kind of computational analysis has become both possible and useful. Structural analysis of DNA-protein interactions can be useful for the classification of DNA-binding proteins [29] and the extraction of general features of the DNA-protein interface [22-24]. Examples of the classification of DNA-binding proteins based on computational structural analysis are shown in section 1.1. This is useful not only for the theoretical understanding of DNA-binding proteins, but also for the computational prediction of DNA-binding sites on DNA and also on protein [78-80].

References

1. Levine M, Tjian R: **Transcription regulation and animal diversity**. *Nature* 2003, **424**(6945):147-151.
2. van Nimwegen E: **Scaling laws in the functional content of genomes**. *Trends Genet* 2003, **19**(9):479-484.
3. Itzkovitz S, Tlusty T, Alon U: **Coding limits on the number of transcription factors**. *BMC Genomics* 2006, **7**:239.
4. Ruvkun G, Hobert O: **The taxonomy of developmental control in Caenorhabditis elegans**. *Science* 1998, **282**(5396):2033-2041.
5. Aoyagi N, Wassarman DA: **Genes encoding Drosophila melanogaster RNA polymerase II general transcription factors: diversity in TFIIA and TFIID components contributes to gene-specific transcriptional regulation**. *J Cell Biol* 2000, **150**(2):F45-50.
6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
7. Latchman DS: **Transcription factors: an overview**. *Int J Biochem Cell Biol* 1997, **29**(12):1305-1312.

8. Stegmaier P, Kel AE, Wingender E: **Systematic DNA-binding domain classification of transcription factors**. *Genome Inform* 2004, **15**(2):276-286.
9. Warnmark A, Treuter E, Wright AP, Gustafsson JA: **Activation functions 1 and 2 of nuclear receptors: molecular strategies for transcriptional activation**. *Mol Endocrinol* 2003, **17**(10):1901-1909.
10. Weigel NL, Moore NL: **Steroid Receptor Phosphorylation: A Key Modulator of Multiple Receptor Functions**. *Mol Endocrinol* 2007.
11. Brivanlou AH, Darnell JE, Jr.: **Signal transduction and the control of gene expression**. *Science* 2002, **295**(5556):813-818.
12. Latchman DS: **Transcription-factor mutations and disease**. *N Engl J Med* 1996, **334**(1):28-33.
13. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: **Molecular Biology of the Cell**, fourth edition edn. Nwe York: Garland Science, a member of the Taylor & Francis Group; 2002.
14. Freidkin I, Katcoff DJ: **Specific distribution of the Saccharomyces cerevisiae linker histone homolog HHO1p in the chromatin**. *Nucleic Acids Res* 2001, **29**(19):4043-4051.
15. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J: **A genomic code for nucleosome positioning**. *Nature* 2006, **442**(7104):772-778.
16. Mishra NC: **Nucleases: Molecular Biology and Applications**. New Jersey: Wiley-Interscience; 2002.
17. Avery OT, MacLeod CM, McCarty M: **Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. 1944**. *Mol Med* 1995, **1**(4):344-365.
18. **Biomedical Hypertexts**
19. Joyce CM, Steitz TA: **Polymerase structures and function: variations on a theme?** *J Bacteriol* 1995, **177**(22):6321-6329.
20. Savarino A: **A historical sketch of the discovery and development of HIV-1 integrase inhibitors**. *Expert Opin Investig Drugs* 2006, **15**(12):1507-1522.

21. Tuteja N, Tuteja R: **Unraveling DNA helicases. Motif, structure, mechanism and function**. *Eur J Biochem* 2004, **271**(10):1849-1863.
22. Jones S, van Heyningen P, Berman HM, Thornton JM: **Protein-DNA interactions: A structural analysis**. *J Mol Biol* 1999, **287**(5):877-896.
23. Lejeune D, Delsaux N, Charloteaux B, Thomas A, Brasseur R: **Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure**. *Proteins* 2005, **61**(2):258-271.
24. Luscombe NM, Laskowski RA, Thornton JM: **Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level**. *Nucleic Acids Res* 2001, **29**(13):2860-2874.
25. Mandel-Gutfreund Y, Schueler O, Margalit H: **Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles**. *J Mol Biol* 1995, **253**(2):370-382.
26. Mirny LA, Gelfand MS: **Structural analysis of conserved base pairs in protein-DNA complexes**. *Nucleic Acids Res* 2002, **30**(7):1704-1711.
27. Nadassy K, Wodak SJ, Janin J: **Structural features of protein-nucleic acid recognition sites**. *Biochemistry* 1999, **38**(7):1999-2017.
28. Pabo CO, Nekludova L: **Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition?** *J Mol Biol* 2000, **301**(3):597-624.
29. Luscombe NM, Austin SE, Berman HM, Thornton JM: **An overview of the structures of protein-DNA complexes**. *Genome Biol* 2000, **1**(1):REVIEWS001.
30. Harrison SC: **A structural taxonomy of DNA-binding domains**. *Nature* 1991, **353**(6346):715-719.
31. Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA: **Mammalian RNA polymerase II core promoters: insights from genome-wide studies**. *Nat Rev Genet* 2007, **8**(6):424-436.
32. Dombroski AJ, Walter WA, Gross CA: **The role of the sigma subunit in promoter recognition by RNA polymerase**. *Cell Mol Biol Res* 1993, **39**(4):311-317.

33. Schultz MC, Reeder RH, Hahn S: **Variants of the TATA-binding protein can distinguish subsets of RNA polymerase I, II, and III promoters**. *Cell* 1992, **69**(4):697-702.

34. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM: **An abundance of bidirectional promoters in the human genome**. *Genome Res* 2004, **14**(1):62-66.

35. Engstrom PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzi L, Tan SL, Yang L *et al*: **Complex Loci in human and mouse genomes**. *PLoS Genet* 2006, **2**(4):e47.

36. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC *et al*: **Genome-wide analysis of mammalian promoter architecture and evolution**. *Nat Genet* 2006, **38**(6):626-635.

37. Schaefer BC: **Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends**. *Anal Biochem* 1995, **227**(2):255-273.

38. Harbers M, Carninci P: **Tag-based approaches for transcriptome research and genome annotation**. *Nat Methods* 2005, **2**(7):495-502.

39. Wold B, Myers RM: **Sequence census methods for functional genomics**. *Nat Methods* 2008, **5**(1):19-21.

40. Sambrook J, Russel DW: **Molecular Cloning: A Laboratory Manual**. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 2001.

41. Fried M, Crothers DM: **Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis**. *Nucleic Acids Res* 1981, **9**(23):6505-6525.

42. Garner MM, Revzin A: **A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system**. *Nucleic Acids Res* 1981, **9**(13):3047-3060.

43. Yang VW: **Eukaryotic transcription factors: identification, characterization and functions**. *J Nutr* 1998, **128**(11):2045-2051.

44. Brenowitz M, Senear DF, Shea MA, Ackers GK: **Quantitative DNase footprint titration: a method for studying protein-DNA interactions**. *Methods Enzymol* 1986, **130**:132-181.
45. Galas DJ, Schmitz A: **DNAse footprinting: a simple method for the detection of protein-DNA binding specificity**. *Nucleic Acids Res* 1978, **5**(9):3157-3170.
46. Brunelle A, Schleif RF: **Missing contact probing of DNA-protein interactions**. *Proc Natl Acad Sci U S A* 1987, **84**(19):6673-6676.
47. Chodosh LA, Carthew RW, Sharp PA: **A single polypeptide possesses the binding and transcription activities of the adenovirus major late transcription factor**. *Mol Cell Biol* 1986, **6**(12):4723-4733.
48. Kwast-Welfeld J, de Belle I, Walker PR, Whitfield JF, Sikorska M: **Identification of a new cAMP response element-binding factor by southwestern blotting**. *J Biol Chem* 1993, **268**(26):19581-19585.
49. Shannon MF, Rao S: **Transcription. Of chips and ChIPs**. *Science* 2002, **296**(5568):666-669.
50. Elnitski L, Jin VX, Farnham PJ, Jones SJ: **Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques**. *Genome Res* 2006, **16**(12):1455-1464.
51. Fredriksson S, Gullberg M, Jarvius J, Olsson C, Pietras K, Gustafsdottir SM, Ostman A, Landegren U: **Protein detection using proximity-dependent DNA ligation assays**. *Nat Biotechnol* 2002, **20**(5):473-477.
52. Rombauts S, Florquin K, Lescot M, Marchal K, Rouze P, van de Peer Y: **Computational approaches to identify promoters and cis-regulatory elements in plant genomes**. *Plant Physiol* 2003, **132**(3):1162-1176.
53. Ohler U, Harbeck S, Niemann H, Noth E, Reese MG: **Interpolated markov chains for eukaryotic promoter recognition**. *Bioinformatics* 1999, **15**(5):362-369.
54. Kondrakhin YV, Kel AE, Kolchanov NA, Romashchenko AG, Milanesi L: **Eukaryotic promoter recognition by binding sites for transcription factors**. *Comput Appl Biosci* 1995, **11**(5):477-488.

55. Bajic VB, Seah SH, Chong A, Zhang G, Koh JL, Brusic V: **Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters**. *Bioinformatics* 2002, **18**(1):198-199.

56. Zhang MQ: **Identification of human gene core promoters in silico**. *Genome Res* 1998, **8**(3):319-326.

57. Prestridge DS: **Predicting Pol II promoter sequences using transcription factor binding sites**. *J Mol Biol* 1995, **249**(5):923-932.

58. Knudsen S: **Promoter2.0: for the recognition of PolII promoter sequences**. *Bioinformatics* 1999, **15**(5):356-361.

59. Reese MG: **Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome**. *Comput Chem* 2001, **26**(1):51-56.

60. Davuluri RV, Grosse I, Zhang MQ: **Computational identification of promoters and first exons in the human genome**. *Nat Genet* 2001, **29**(4):412-417.

61. Blanchette M, Sinha S: **Separating real motifs from their artifacts**. *Bioinformatics* 2001, **17 Suppl 1**:S30-38.

62. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCH: A tool for searching transcription factor binding sites in DNA sequences**. *Nucleic Acids Res* 2003, **31**(13):3576-3579.

63. Sandelin A, Wasserman WW, Lenhard B: **ConSite: web-based prediction of regulatory elements using cross-species comparison**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W249-252.

64. Marinescu VD, Kohane IS, Riva A: **MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes**. *BMC Bioinformatics* 2005, **6**:79.

65. Loots GG, Ovcharenko I: **rVISTA 2.0: evolutionary analysis of transcription factor binding sites**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W217-221.

66. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment**. *Science* 1993, **262**(5131):208-214.

67. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.

68. Liu X, Brutlag DL, Liu JS: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes**. *Pac Symp Biocomput* 2001:127-138.

69. Sinha S, Tompa M: **YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation**. *Nucleic Acids Res* 2003, **31**(13):3586-3588.

70. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements**. *Nat Rev Genet* 2004, **5**(4):276-287.

71. GuhaThakurta D: **Computational identification of transcriptional regulatory elements in DNA sequence**. *Nucleic Acids Res* 2006, **34**(12):3585-3598.

72. Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I: **Identification of transcription factor binding sites with variable-order Bayesian networks**. *Bioinformatics* 2005, **21**(11):2657-2666.

73. Huang W, Umbach DM, Ohler U, Li L: **Optimized mixed Markov models for motif identification**. *BMC Bioinformatics* 2006, **7**:279.

74. Zhao X, Huang H, Speed TP: **Finding short DNA motifs using permuted Markov models**. *J Comput Biol* 2005, **12**(6):894-906.

75. Naughton BT, Fratkin E, Batzoglou S, Brutlag DL: **A graph-based motif detection algorithm models complex nucleotide dependencies in transcription factor binding sites**. *Nucleic Acids Res* 2006, **34**(20):5730-5739.

76. Zhang MQ, Marr TG: **A weight array method for splicing signal analysis**. *Comput Appl Biosci* 1993, **9**(5):499-509.

77. King OD, Roth FP: **A non-parametric model for transcription factor binding sites**. *Nucleic Acids Res* 2003, **31**(19):e116.

78. Narlikar L, Gordan R, Ohler U, Hartemink AJ: **Informative priors based on transcription factor structural class improve de novo motif discovery**. *Bioinformatics* 2006, **22**(14):e384-392.

79. Sandelin A, Wasserman WW: **Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics**. *J Mol Biol* 2004, **338**(2):207-215.
80. Morozov AV, Siggia ED: **Connecting protein structure with predictions of regulatory sites**. *Proc Natl Acad Sci U S A* 2007, **104**(17):7068-7073.
81. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW: **Identification of conserved regulatory elements by comparative genome analysis**. *J Biol* 2003, **2**(2):13.
82. Siddharthan R, Siggia ED, van Nimwegen E: **PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny**. *PLoS Comput Biol* 2005, **1**(7):e67.
83. Ureta-Vidal A, Ettwiller L, Birney E: **Comparative genomics: genome-wide analysis in metazoan eukaryotes**. *Nat Rev Genet* 2003, **4**(4):251-262.
84. Narlikar L, Gordan R, Hartemink AJ: **Nucleosome Occupancy Information Improves de novo Motif Discovery**. In: *RECOMB: 2007*: Springer; 2007: 107-121.
85. Tomovic A, Oakeley EJ: **Position dependencies in transcription factor binding sites**. *Bioinformatics* 2007, **23**(8):933-941.
86. Frith MC, Hansen U, Weng Z: **Detection of cis-element clusters in higher eukaryotic DNA**. *Bioinformatics* 2001, **17**(10):878-889.
87. GuhaThakurta D, Stormo GD: **Identifying target sites for cooperatively binding factors**. *Bioinformatics* 2001, **17**(7):608-621.
88. Frith MC, Li MC, Weng Z: **Cluster-Buster: Finding dense clusters of motifs in DNA sequences**. *Nucleic Acids Res* 2003, **31**(13):3666-3668.
89. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.

2. Position dependencies in transcription factor binding sites (paper I)

Most of the available tools for transcription factor binding site prediction are based on methods which assume no sequence dependence between the binding site base positions. The primary objective of this work was to investigate the statistical basis for either a claim of dependence or independence, to determine whether such a claim is generally true, and to use the resulting data to develop improved scoring functions for binding-site prediction. Using three statistical tests, the number of binding sites showing dependent positions has been analyzed. Transcription factor-DNA crystal structures are also analysed, in order to find a possible biological explanation of dependent positions. The final conclusions were that some factors show evidence of dependencies, whereas others do not. It was observed that the conformational energy (Z-score) of the transcription factor-DNA complexes was lower (better) for sequences that showed dependency than for those that did not ($P < 0.02$). It can be suggested that where evidence exists for dependencies, these should be modelled to improve binding-site predictions. However, when no significant dependency is found, this correction should be omitted. This may be done by converting any existing scoring function which assumes independence into a form which includes a dependency correction. An example of such an algorithm and its implementation as a web tool is presented.

All supplemental materials for this paper are available in this chapter, and implementation of the presented algorithm is publicly available from http://promoterplot.fmi.ch/cgi-bin/dep.html

BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 8 2007, pages 933–941
doi:10.1093/bioinformatics/btm055

Sequence analysis

Position dependencies in transcription factor binding sites

Andrija Tomovic and Edward J. Oakeley*

Friedrich Miescher Institute for Biomedical Research, Novartis Research Foundation, Maulbeerstrasse 66, CH-4058 Basel, Switzerland

Received on October 29, 2006; revised on January 17, 2007; accepted on February 9, 2007

Advance Access publication February 18, 2007

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Most of the available tools for transcription factor binding site prediction are based on methods which assume no sequence dependence between the binding site base positions. Our primary objective was to investigate the statistical basis for either a claim of dependence or independence, to determine whether such a claim is generally true, and to use the resulting data to develop improved scoring functions for binding-site prediction.

Results: Using three statistical tests, we analyzed the number of binding sites showing dependent positions. We analyzed transcription factor–DNA crystal structures for evidence of position dependence. Our final conclusions were that some factors show evidence of dependencies whereas others do not. We observed that the conformational energy (Z-score) of the transcription factor–DNA complexes was lower (better) for sequences that showed dependency than for those that did not ($P < 0.02$). We suggest that where evidence exists for dependencies, these should be modeled to improve binding-site predictions. However, when no significant dependency is found, this correction should be omitted. This may be done by converting any existing scoring function which assumes independence into a form which includes a dependency correction. We present an example of such an algorithm and its implementation as a web tool.

Availability: http://promoterplot.fmi.ch/cgi-bin/dep.html

Contact: edward.oakeley@fmi.ch

Supplementary information: Supplementary data (1, 2, 3, 4, 5, 6, 7 and 8) are available at *Bioinformatics* online.

1 INTRODUCTION

The transcription of genes is controlled by transcription factor proteins (TFs) which bind to short DNA sequences known as transcription factor binding sites (also known as DNA-binding motifs or *cis*-regulatory sequences). TF-binding sites are usually very short and highly degenerate, and such short sequences are expected to occur at random every few hundred base pairs. This makes their prediction extremely difficult. An important task in the computational prediction of TF-binding sites is reducing the false positive rate while still retaining a high sensitivity. Currently, predictions rely on either scanning or *ab initio* methods. Scanning methods infer binding sites from known, experimentally verified binding sequences. Example tools include ConSite (Sandelin *et al.*, 2004a), Match (Kel *et al.*, 2003), Mapper (Marinescu *et al.*, 2005), Patser (Hertz *et al.*, 1990), and rVista (Loots and Ovcharenko, 2004; Loots *et al.*, 2002). *Ab initio* approaches infer specificities without any prior knowledge of binding sites, based on sequence homology. Example tools include Gibbs sampler (Lawrence *et al.*, 1993), MEME (Bailey and Elkan, 1994), Bioprospector (Liu *et al.*, 2001), Yeast motif finder (Sinha and Tompa, 2003) and ANN-Spec (Workman and Stormo, 2000). Until recently, the most popular way of modeling binding sites was to assume that each base in the site occurs independently, e.g. consensus sequence (Day and McMorris, 1992), matrix profiles (Stormo *et al.*, 1982) and sequence logos (Schneider and Stephens, 1990); for a review see (Wasserman and Sandelin, 2004). Methods based on the assumption of independence between positions are simple with small numbers of parameters, making them easy to implement. These methods are widely used and often considered as acceptable models for binding-site predictions (Benos *et al.*, 2002a). However, recent experimental evidence (Benos *et al.*, 2002b; Bulyk *et al.*, 2002; Man and Stormo, 2001; Udalova *et al.*, 2002; Wolfe *et al.*, 1999) has prompted the development of models which incorporate position dependencies. The related methods include Bayesian networks (Barash, 2003), permuted Markov models (Zhao *et al.*, 2005), Markov chain optimization (Ellrott *et al.*, 2002), hidden Markov models (Marinescu *et al.*, 2005), non-parametric models (King and Roth, 2003), and generalized weight matrix models (Zhou and Liu, 2004). Methods based on position-dependency models usually have better binding site prediction accuracy with lower false positive rates. But these methods require more complicated mathematical tools, with more parameters to estimate, and require more experimental data than are typically available (Barash, 2003; Ellrott *et al.*, 2002; King and Roth, 2003; Marinescu *et al.*, 2005; Zhao *et al.*, 2005; Zhou and Liu, 2004). The purpose of this work is to investigate whether or not TFs show position dependencies in their binding sites. We suggest a rigorous statistical approach for testing dependencies. Our findings indicate that there is no universal answer. Some factors seem to show dependencies whereas others do not. We, therefore, decided to allow both possibilities within our model. Our method for modeling dependencies is simply an extension of methods which assume position independencies. It does not require complex

*To whom correspondence should be addressed.

mathematical tools or training data sets (and thus more data), and we will show that it performs much better than existing tools when dependencies are found and no worse when they are not. We also analyzed available structural data to see if any of the observed position dependencies can be explained by 3D structures. We found that dependencies may be partially explained by the 3D structure of TF–DNA complexes. TFs with dependent positions also appear to fit their target sequences better than those without dependencies.

2 METHODS

2.1 Testing dependencies

In this section, we describe methods to test dependencies in binding sites.

Let us suppose that we have n binding sites of length k for a given TF:

$$\begin{matrix} b_1^1 & b_2^1 & \dots & b_k^1 \\ \dots & & & \\ b_1^n & b_2^n & \dots & b_k^n \end{matrix} \tag{1}$$

where $b_i^j \in \{a,c,g,t\}$, and $1 \le i \le k$, and $1 \le j \le n$. We introduce the notation: B_i and B_j to represent random variables which can take values from the set $\{a,c,g,t\}$, indices i and j represent positions in the binding sites and $1 \le i, j \le k$ and $i \ne j$,

$$B_i: \begin{pmatrix} a & c & g & t \\ P(a,i) & P(c,i) & P(g,i) & P(t,i) \end{pmatrix} \tag{2}$$

and likewise for B_j.

Let $N(i)$ be a vector of the frequencies $N(i) = [N(a,i), N(c,i), N(g,i), N(t,i)]$ where, $N(a,i)$ is the frequency of base a at position i and so on. Similarly, for column j we introduce a frequency vector $N(j)$. Using a maximum likelihood approach and the method of Lagrange multipliers, we can estimate probabilities:

$$P(b,i) = \frac{N(b,i)}{n}, \quad P(b,j) = \frac{N(b,j)}{n} \tag{3}$$

where b is one of the bases $\{a,c,g,t\}$.

First, we can calculate mutual information (Chiu and Kolodziejczak, 1991), a quantitative measure of pairwise sequence covariation. The mutual information between positions i and j is given by:

$$M_{ij} = \sum_{b_i,b_j} P(b_i,b_j,i,j) \log_2 \frac{P(b_i,b_j,i,j)}{P(b_i,i)P(b_j,j)} \tag{4}$$

where, the probability $P(b_i,b_j,i,j)$ can be estimated using the maximum-likelihood method and the method of Lagrange multipliers:

$$P(b_i,b_j,i,j) = \frac{N(b_i,b_j,i,j)}{n} \tag{5}$$

where, $N(b_i,b_j,i,j)$ is the frequency of base pairs b_ib_j at positions i and j. This is a descriptive measure of divergence from independence of i and j. M_{ij} varies between 0 and 2 bits. It is maximal when i and j are perfectly correlated. If i and j are uncorrelated, the mutual information is zero. Very often we do not have extreme values of M_{ij}, and we cannot deduce if i and j are independent using only the value of M_{ij}. In order to identify positions that may not be highly correlated as measured by M_{ij}, but are as correlated as they can be given the limited variability of the individual positions, we can calculate two other values (Gutell *et al.*, 1992):

$$R_1(i,j) = \frac{M_{ij}}{H_i}, \quad R_2(i,j) = \frac{M_{ij}}{H_j} \tag{6}$$

where H_i and H_j are entropies for positions i and j, respectively.

$$H_i = -\sum_b P(b,i)\log_2 P(b,i), \quad H_j = -\sum_b P(b,j)\log_2 P(b,j) \tag{7}$$

Both R values vary between 0 and 1 and, in general, they are not equal. Therefore, if we use only M_{ij} we may miss some correlated positions, but some of these may be detected using R-values. However, it should be emphasized that we cannot easily estimate the significance of R-values. So, we will have false positives as well as true correlated positions. R-values are also descriptive measures of dependencies between two positions. A more formal way to test dependencies is hypothesis testing:

$$\begin{aligned} &H_0\text{: positions } i \text{ and } j \text{ are independent} \\ &H_1\text{: otherwise.} \end{aligned} \tag{8}$$

To test this hypothesis, we can use a χ^2-test of independence (Ellrott *et al.*, 2002) on each pair of positions i and j:

$$X^2 = \sum_{b_i,b_j} \frac{(P(b_i,b_j,i,j) - P(b_i,i)P(b_j,j))^2}{P(b_i,i)P(b_j,j)} \tag{9}$$

The distribution of X^2 statistics is close to a χ^2 distribution with $(|b_i|-1)\times(|b_j|-1)$ degrees of freedom, where $|b_i|$ is the number of bases for which $P(b_i,i)$ is not zero, and likewise for $|b_j|$. So, using X^2 statistics and χ^2 distributions we can test the hypothesis at a given significance level e.g. 0.05. This hypothesis may also be tested using a G-test of independence (log-likelihood ratio test) (Sokal and Rohlf, 2003). For each pair of positions i and j, we can calculate G statistics:

$$G = 2\sum_{b_i,b_j} P(b_i,b_j,i,j) \ln\left(\frac{P(b_i,b_j,i,j)}{P(b_i,i)P(b_j,j)}\right) \tag{10}$$

The distribution of G statistics is close to χ^2 with $(|b_i|-1)\times(|b_j|-1)$ degrees of freedom where $|b_i|$ is the number of bases for which $P(b_i,i)$ is not zero, and likewise for $|b_j|$. M_{ij} corresponds to a G-statistics value if we log transform it. A general problem with both χ^2 and G-tests is small sample sizes, i.e. small expected frequencies (in our notation these are the values $nP(b_i,i)$ and $nP(b_j,j)$). This is because the number of known binding sites is usually small. Cochran (Cochran, 1954) suggested that independence may be tested so long as we have more than one degree of freedom. A minimum expected value of 1 is allowed, provided that no more than 20% of the categories have expected values below 5. Here, X^2 statistics have been shown to be valid with fewer samples and more sparse tables than G statistics. The G-statistic distribution is usually a poor approximation to χ^2 when expected frequencies are <5 (Agresti, 1990; Koehler, 1986; Koehler and Larntz, 1980; Larntz, 1978). William's correction for G (Williams, 1976) partially addresses this:

$$G_{\text{adj}} = \frac{G}{q}, \quad q = 1 + \frac{(a^2-1)}{6nv} \tag{11}$$

where, $a = (|b_i|-1)\times(|b_j|-1) - 1$, and $v = a - 1$ as this provides a better approximation to the χ^2 distribution. Conahan found that if expected frequencies are higher than 10, G statistics approximate well to the exact multinomial probability distribution (Conahan, 1970). She found that G statistics were adequate and better than X^2 statistics, where there are more than five degrees of freedom and expected frequencies greater than or equal to 3. In all other cases she recommends the exact test. Larantz, in his comparison of G and X^2 statistics, did not consider the corrections of G statistics when drawing his conclusion that X^2 statistics fits the theoretical chi-squared distribution better than G statistics do (Larntz, 1978). Sokal *et al.* (Sokal and Rohlf, 2003) showed that G statistics with William's correction approximates to the χ^2 distribution more closely than they do without the correction. It is very difficult to find a single rule to cover all cases when the observed distributions of G statistics and X^2 statistics are close to real χ^2 distributions, if we have small expected

frequencies (Agresti, 1990). A safer way to test the hypothesis of dependence is, therefore, to use exact methods like the exact randomization (nonparametric) test (Sokal and Rohlf, 2003). The problem with this test is that, even though we have small sample numbers, there are a large number of possible outcomes, and their complete enumeration is impractical. Because of this, we have to use a Monte Carlo randomization test (Davison and Hinkley, 1997; Manly, 1997), in which the problem is solved by random sampling from a simulated population. Monte Carlo randomization tests can be performed using X^2 or G statistics. We used X^2 statistics with 10 000 replications in the statistics package R (GNU software).

Two random variables b_i and b_j are independent if

$$P(B_i, B_j) = P(B_i)P(B_j). \tag{12}$$

Thus we can test the following hypotheses for dependence/independence (instead of hypothesis testing (9)):

$$\begin{aligned} &H_0\text{: distributions } P(B_i, B_j) \text{ and } P(B_i)P(B_j) \text{ are the same} \\ &H_1\text{: otherwise.} \end{aligned} \tag{13}$$

This form of hypothesis testing corresponds to a multinomial goodness-of-fit test. As in (Bejerano, 2003, 2006; Bejerano *et al.*, 2004), we can test for a correlation between TF-binding site positions using exact P-values (for hypothesis testing (13)). This approach gives more accurate results than either χ^2 or G-tests (Bejerano, 2003, 2006; Bejerano *et al.*, 2004). The only problem with this approach is that it is computationally expensive. However, a recent publication (Keich and Nagarajan, 2006) has shown that grid approximations yield almost identical results for the P-values but in far less time (Bejerano, 2006). The final method we have used to test dependencies is a Bayesian approach (Minka, 2003; Zhou and Liu, 2004). We can calculate the Bayes factor $BF(H_0; H_1)$ for hypothesis testing as follows (full derivation of formula (4) can be found in Supplemental Material 1—derivation 1)

$$\begin{aligned} BF(H_0; H_1) = &\frac{\Gamma\left(\sum_{b_i, b_j} \alpha_{b_i b_j}\right)}{\Gamma\left(n + \sum_{b_i, b_j} \alpha_{b_i b_j}\right)} \prod_{b_i} \frac{\Gamma\left(N(b_i, i) + \alpha_{b_i}\right)}{\Gamma\left(\alpha_{b_i}\right)} \\ &* \prod_{b_j} \frac{\Gamma\left(N(b_j, j) + \alpha_{b_j}\right)}{\Gamma\left(\alpha_{b_j}\right)} \prod_{b_i, b_j} \frac{\Gamma\left(\alpha_{b_i b_j}\right)}{\Gamma\left(N(b_i, b_j, i, j) + \alpha_{b_i b_j}\right)} \end{aligned} \tag{14}$$

We choose $\alpha_{b_i b_j} = 1$ and $\alpha_{b_i} = \sum_{b_j} \alpha_{b_i b_j}$ and the calculation should include only bases b_i, b_j for which $N(b_i, i) \neq 0$ and $N(b_j, j) \neq 0$.

Using Stirling's approximation ($\log\Gamma(x+1) \approx x \log x - x$) it can be shown that (Supplemental Material 1—derivation 2)

$$\log_2(BF(H_0; H_1)) \approx -nM_{ij} \tag{15}$$

This gives us the relationship between BF and mutual information (Minka, 2003). The relationship between these two values is better when the sample size n is higher (due to the use of Stirling's approximation). We used formula (20) to calculate BF, and report that when $BF(H_0; H_1) < 0.1$ there is strong evidence against the null hypothesis.

Thus, in this article we used three distinct methods for dependence testing between the TF site base positions. These methods were:

(i) Monte Carlo randomization test with X^2 or G statistics
(ii) Exact multinomial goodness-of-fit test
(iii) Bayesian hypothesis testing.

There is always a danger of type I errors (rejecting the null hypothesis when in fact it is true) when applying multiple tests to data. These may be minimized with Bonferroni's correction or its extensions/variants (e.g. Dunn–Šidák, Holm's, Simes–Hochberg or Hommel's method). The Bonferroni adjustment of P-value ($0.05/k$, where k is the number of tests) is very stringent and can introduce type II errors, which are also important. The use of Bonferroni is much debated (Perneger, 1998).

As a compromise, in the case of the Bayesian test, we propose that a more stringent BF factor $BF(H_0; H_1) < 0.1$ could be used to report stronger evidence against the null hypothesis.

2.2 New scoring function

Any existing scoring function which works with models that assume independence between positions within binding sites, can easily be modified to incorporate dependencies. These new functions do not have dramatically more parameters, and do not require additional data or complex mathematical approaches.

If we have n binding sites of length k for a given TF and sequence l with length k, then to determine if a putative-binding site is for a given TF we will follow the notation of (Wasserman and Sandelin, 2004) where, $w_{b,i}$ is a position weight matrix (PWM) value of base b in position i, calculated by:

$$W_{b,i} = \log_2 \frac{P(b, i)}{P(b)} \tag{16}$$

where $P(b)$ is the background probability of base b ($P(b) = 0.25$) and $P(b, i)$ is a corrected probability of base b at position i, and is calculated by:

$$P(b, i) = \frac{N(b, i)}{n} + a(b) \tag{17}$$

where $a(b)$ is smoothing parameter ($a(b) = 0.01$).

The fit of any given DNA sequence can be quantitatively scored by summing all the values of $W_{b,i}$ for every base in the sequence (hereafter, we will refer to this 'old' scoring function as S_{old}):

$$S_{\text{old}} = \sum_{i=l}^{k} w_{l_i, i} \tag{18}$$

For a large set of well-characterized binding sites, these scores are proportional to the factor-binding energies (Stormo, 2000).

To incorporate position dependencies, we will extend this function and this model for the representation of the TF-binding sites in the following way.

First, we will introduce a corrected probability for the bases $b_1 b_2 \ldots b_m$ in $i_1 i_2 \ldots i_m$ dependent positions.

$$P(b_1, \ldots, b_m, i_1, \ldots, i_m) = \frac{N(b_1, \ldots, b_m, i_1, \ldots, i_m)}{n} + a(b_1, \ldots, b_m) \tag{19}$$

$a(b_1, \ldots, b_m)$ is a smoothing parameter and can be calculated by:

$$a(b_1, b_2, \ldots, b_m) = a(b_1) \ldots a(b_m) \tag{20}$$

Then we can calculate values which correspond to PWM values:

$$W_{b_1, \ldots, b_m, i_1, \ldots, i_m} = \log_2 \frac{P(b_1, \ldots, b_m, i_1, \ldots, i_m)}{P(b_1) \ldots P(b_m)} \tag{21}$$

Finally, the new scoring function (S_{new}), which incorporates dependencies, can be expressed thus:

$$\begin{aligned} S_{\text{new}} = &\sum_{i=l}^{k_1} W_{l_i, i} + \sum_{i=1}^{k_2} W_{l_{j_i}, l_{j_{i+1}}, j_i, j_{i+1}} + \cdots + \\ &+ \sum_{i=l}^{k_m} W_{l_{j_i}, \ldots, l_{j_{i+m-1}}, j_i, \ldots, j_{i+m-1}} \end{aligned} \tag{22}$$

where, k_1 is the number of independent positions, k_2 is the number of dependent positions order 2 (nucleotides at positions j_i and j_{i+1}) and k_m the number of dependent positions order m (nucleotides at positions j_i, $j_{i+1}, \ldots, j_{i+m-1}$). Higher-order dependencies can be constructed from

the second-order dependencies in the following way: if we analyze three positions i_1, i_2 and i_3, and if every two combinations ($i_1 - i_2$, $i_1 - i_3$ and $i_2 - i_3$) are dependent, then we can claim that positions i_1, i_2 and i_3 show third-order dependencies. This approach may be extended to find mth-order dependencies between k_m positions of a binding site. For the new scoring function (22), higher order dependencies can be constructed in a less stringent way: if we find when analyzing three positions i_1, i_2 and i_3 that only two combinations ($i_1 - i_2$, $i_2 - i_3$ or $i_1 - i_3$) are dependent, we can say that there are third order dependencies among positions i_1, i_2 and i_3. This will not have any influence on the final results (because of equation (12)) and the logarithm property ($\log(P(B_i, B_j))$ will tend towards $\log(P(B_i)) + \log(P(B_j))$). Small differences may be observed because of the smoothing parameters, but this helps in the practical implementation of new scoring function.

Binding scores calculated by the scoring functions S_{old} and S_{new} can be normalized according to (Bucher, 1990; Tsunoda and Takagi, 1999):

$$S'_{\text{old}} = \frac{S_{\text{old}} - S^{\min}_{\text{old}}}{S^{\max}_{\text{old}} - S^{\min}_{\text{old}}}, \quad S'_{\text{new}} = \frac{S_{\text{new}} - S^{\min}_{\text{new}}}{S^{\max}_{\text{new}} - S^{\min}_{\text{new}}} \tag{23}$$

where $S^{\min}_{\text{old}}$, $S^{\max}_{\text{old}}$ are the hypothetical minimum and maximum for S_{old} and $S^{\min}_{\text{new}}$, $S^{\max}_{\text{new}}$ are the hypothetical minimum and maximum for S_{new} (analytic formula for their calculation is given in Supplemental Material 1).

For the final implementation of the function (22), it is useful to construct sequence dependency corrected matrices of TFs. However, in practice, this can be very inefficient because the dimensions of these matrices can be very high with a lot of zeros. Because of this, we provide a database (available at http://www.fmi.ch/members/andrija.tomovic/database.txt) with sequences and dependent positions written below (estimated using a Monte Carlo randomization test with X^2 without Bonferroni's correction or exact multinomial goodness-of-fit without Bonferroni's correction or Bayesian hypothesis testing with $BF(H_0; H_1) < 0.1$ and higher order of dependencies in less stringent variant). This is a compact and readable format of sequence dependency corrected matrices of TFs from the JASPAR database (Lenhard and Wasserman, 2002; Sandelin *et al.*, 2004b). For the identification of TF-binding sites by scoring function (22), we suggest using the all-atom model, like it is used with function (18). In combination with databases of good quality binding sites (such as JASPAR) all-atom methods give better accuracy. If we cut the length of binding sites, there may be dependent positions in this region which will be lost to our function (22). Both the new (22) and old (18) scoring functions are linear in complexity, so cutting would not improve performance much.

3 RESULTS AND DISCUSSION

3.1 Distributions of transcription factors with dependent positions

To determine the distributions of TFs with dependent positions, we used the public database JASPAR (Lenhard and Wasserman, 2002; Sandelin *et al.*, 2004b) which contains experimentally determined, high-quality binding sites. The JASPAR database represents a curated and non-redundat data-set (Lenhard and Wasserman, 2002; Sandelin *et al.*, 2004b). We selected all TFs for which there were binding sequences (not only matrix profiles) and the final data set contained 107 TFs with 3239 binding sites. We applied three different tests (Section 2.1) to each of these binding sites to establish how many factors showed position dependencies (Table 1). We also show the effect of either applying Bonferroni's corrections or using the more stringent $BF(H_0; H_1) < 0.1$ cut off. Rows A, B and C of Table 1 may include some false positives, but rows D, E and F have a false negative problem. A complete list of every pair of positions for each TF is given in Supplemental Material 2. We also report values of M_{ij}, R_1 and R_2, as well as G-statistic values with their degrees of freedom and P-values. In addition, we report the adjusted G-statistic values with their degrees of freedom and adjusted G-test P-values; the X^2 statistics together with their degrees of freedom and P-values; and also the average value of expected frequencies and the percentage of expected values smaller than 5 and smaller than 3. Finally, in this table we report the P-values of the Monte Carlo randomization test with X^2 statistics, the exact multinomial 'goodness-of-fit' test and the Bayesian factor (BF) values. From this analysis, we observe that the sample sizes are not appropriate for either chi-squared or G-tests of independence (column H in Supplemental Material 2). As discussed previously (Section 2.1), this implies that these two tests will give poor probability estimates. The values of M_{ij}, R_1 and R_2 may be used as descriptive measures of position associations. There is good agreement between results produced using the three 'statistically correct' tests we attempted. The most stringent is the exact multinomial goodness-of-fit test, and the least stringent is the Monte Carlo randomization test. Almost every pair of dependent positions predicted by the exact multinomial goodness-of-fit test is also reported by the other two tests. The Monte Carlo randomization test gives more precise probabilities than either the chi-squared or G-tests, but with low power because of the lack of experimental data (small sample size).

Table 1. Distributions of TFs with dependent positions tested

Statistical test	TFs with dependent positions
A	74.77%
B	49.52%
C	62.62%
D	38.32%
E	23.26%
F	26.17%

A—Monte Carlo randomization test without Bonferroni's correction; B—Exact multinomial 'goodness-of-fit' test without Bonferroni's correction; C—Bayesian hypothesis testing $BF(H_0; H_1) < 0.1$; D—Monte Carlo randomization test with Bonferroni's correction; E—Exact multinomial 'goodness-of-fit' test with Bonferroni's correction; F—Bayesian hypothesis testing $BF(H_0; H_1) < 0.01$.

In addition, we looked to see if the length and number of known binding sites were different between the groups of TFs with and without dependent positions (Table 2). The variances of these two groups are not statistically different (tested by Bartlett's test). Using Student's t-test, we tested the null hypothesis that mean length and number of binding sites between the two groups are equal against a one-tailed alternative hypothesis that TFs without dependent positions have shorter lengths and smaller numbers of known binding sites. In each case, we obtained P-values less than 0.05 and thus we should reject the null hypothesis and accept the alternative.

Table 2. Average length and number of binding sites between a group of TFs with dependent positions and a group of TFs without dependent positions

Statistical test	Average length of TFs binding sites		Average number of known binding sites	
	I	II	I	II
A	11.67	8.25	32.85	22.64
B	12.15	9.43	34.66	25.77
C	11.66	9.3	35.791	20.775
D	12.19	9.89	39.15	24.61
E	11.92	10.265	45.04	25.82
F	12.00	10.34	50.96	22.91

I—group with dependent positions; II—group without dependent positions; A, B, C, D, E, F — notation the same as in Table 1.

These results imply that more factors may show dependencies once additional binding-site data becomes available.

Based on the second-order dependencies (dinucleotide dependencies), it is possible to construct higher order dependencies, as explained in section 2. It is clear that when we have such dependencies, there are lower order dependencies in all combinations. Because of this, it is useful to analyze distributions of dependencies of different orders k_m ($2 \leq k_m \leq 9$) constructed in a more stringent way (Supplemental Material 3). We analyzed the distributions of TFs with dependent positions in structural classes of TF–DNA-binding domains. We wanted to investigate whether there is any tendency for certain folds to have position dependencies (Supplemental Material 4). We noticed that some structural classes contain TFs with position dependencies in their binding sites detected by almost all statistical tests, such as: T-BOX, P53, AP2, TRP, CAAT-box and MADS. Other classes contain TFs without dependent positions like: ZH-FINGER-DOF, ZH-FINGER-GATA, HOMEO/CAAT and 'Unknown' class. However, the major structural classes contain TFs with and without dependent positions (bZIP, nuclear receptor, etc.).

3.2 Do position dependencies relate to 3D structures?

We wanted to investigate possible biological explanations of the dependent positions we predicted. We investigated this by examining 3D crystal structures when available. Possible explanations of dependency include:

- active amino acids might interact with dependent nucleotides either singly or in pairs via hydrogen bonds or salt bridges;
- conformational changes in the structure of DNA caused by one dependent base may alter the accessibility of the other dependent bases to the binding site;
- something else.

We selected 32 TF–DNA co-crystal pairs of structures from the PDB database at resolutions better than 3.0 Å (Berman *et al*., 2000) corresponding to TFs with published binding sites in JASPAR (September 2006) (Table Sup3-1 in Supplemental Material 3). Direct contacts between bases and amino acids were investigated (Table Sup3-2 in Supplemental Material 3).

There is no clear one-to-one correspondence between dependent DNA-binding positions and their interactions with TF. This is not a big surprise because these proteins recognize specific DNA sequences not only via direct contact but also indirectly, through specific sequence-dependent DNA conformations, distortions or water-mediated contacts (Sarai and Kono, 2005). Amino acids neighboring dependent bases may be different from those around independent positions. In addition, mutations in bases which do not directly contact the amino acid may still affect the binding affinity (see references listed in Sarai and Kono, 2005).

Next, we wanted to check whether there were any relationships between dependent positions and conformational changes of the DNA. We could calculate structural parameters to describe the 3D nucleic acid structures using the software package 3DNA (Lu and Olson, 2003), but there are many parameters (shift, slide, rise, tilt, roll and twist) to describe the structure of DNA, and because we have relatively few sequences in our data set it is difficult to identify significant effects. Similarly, if we want to investigate spatial distribution patterns of neighboring amino acids around dependent positions, we will have a data-mining problem.

We decided to use the energy Z-scores (Ahmad *et al*., 2006; Gromiha *et al*., 2004; Kono and Sarai, 1999) for TF–DNA complexes for both 'direct' and 'indirect' readouts. The energy Z-score for direct readouts quantifies the spatial distributions of side chains around base pairs, and represents the base–amino acid interaction energy. The energy Z-score for indirect readouts quantifies DNA conformation, and represents the conformational energy of DNA. The more negative the Z-score, the better the target sequence fits into a given structure (Ahmad *et al*., 2006). The list of all Z-score values can be found in Supplemental Material 4. We tested the Z-scores using a one-tailed Student's *t*-test (Table 3). The direct readout showed no difference between TFs with dependent or independent positions ($P > 0.1$). However, the conformational energy (indirect readout) was always significantly lower for TFs with dependent positions ($P < 0.02$). This means that TFs with dependent positions fit their target DNA motifs better than those without. These results suggest a possible relationship between position dependencies and the 3D structure of TFs.

Table 3. Average Z-score for direct and indirect readout for: **I**—a group of TFs with dependent positions; and **II**—a group of TFs without dependent positions

Statistical test	Average Z-score (direct readout)			Average Z-score (indirect readout)		
	I	II	*P*-value	I	II	*P*-value
A	−2.5	−2.62	–	−2.8	−1.791	0.00565**
B	−2.67	−2.42	0.383	−3.0914	−2.01	0.0016**
C	−2.667	−2.25	0.31	−2.747	−1.907	0.02*
D	−3.054	−2.26	0.17	−3.22	−2.09	0.00152**
E	−3.44	−2.3	0.111	−3.33	−2.29	0.0147*
F	−3.1025	−2.32	0.186	−3.497	−2.147	0.0005***

*$P<0.05$, **$P<0.01$, ***$P<0.001$.
A, B, C, D, E, F—notation the same as in Table 1. The variances of groups I and II are not statistically different (Bartlett's test).

We investigated if DNA sequence length influences the conformational energy. In the 32 cases we studied where we have both a 3D crystal structure and a JASPAR matrix ID, we performed one- and two-tailed *t*-tests on the lengths of sequences found to show dependencies and those without dependencies from each of the six dependency tests we investigated. These results showed that five out of six of the tests (two-tailed) and three out of six (one-tailed) do not show significant differences in sequence length between the two groups for which we have conformational energies (Supplemental Material 6). If the conformation of the DNA fragment is not sequence specific, then the conformational energy is expected to fluctuate independently of fragment size. But, if the conformation is sequence specific, then the total energy should decrease with the size although the average energy per base will not decrease if the energy distribution is uniform (A. Sarai, personal communication). For these reasons, we believe that sequence length is not the major factor contributing to the significantly lower conformational energies we found for the group of TFs with dependent positions.

We analyzed relationships between dependent position and DNA stiffness to show the influence of DNA stiffness on protein–DNA binding specificity (Gromiha, 2005). We calculated the average stiffness of DNA using the structure-based sequence-dependent stiffness scale (Gromiha, 2005) for binding sites with and without position dependencies (Supplemental Material 7). In two cases, we found that the average stiffness values are significantly larger (one-tailed Student's *t*-test $P<0.028$) for sites with dependent positions (detected by Bayesian hypothesis testing in both variants) than without dependent positions. However, in the other four cases no significant differences were found.

3.3 Evaluation of a new scoring function for the prediction of TF-binding sites

The evaluation of *ab initio* methods for the prediction of TF-binding sites is described in (Tompa *et al.*, 2005). Here, we will perform a slightly different validation. In order to evaluate the new scoring function given by (22) and (23), we performed a validation using both synthetic and experimentally verified data.

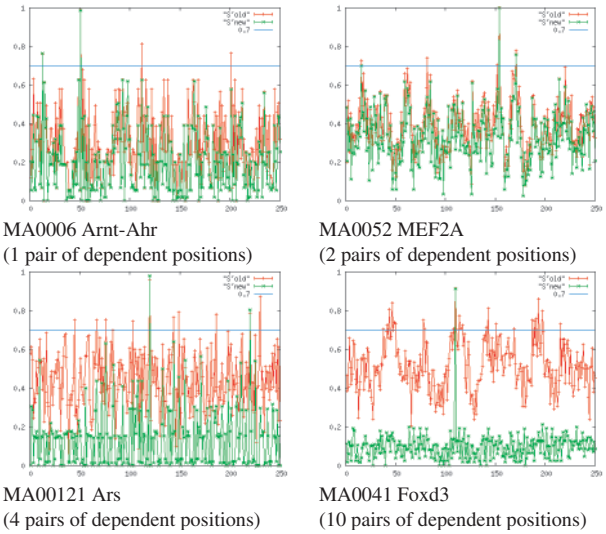


Fig. 1. Comparison of old and new scoring functions with synthetic data.

First, we generated a random sequence from a third-order Markov model background distribution using the program RSA (van Helden, 2003). In this sequence, we planted binding site 9 of the TF MA0006 at position 51. We had found one dependent position in this TF. We then calculated a normalized scoring value for each position in the sequence, using both the old and new functions. We assigned a threshold of 0.7 as indicating a match for a binding site (Fig. 1). The new scoring function made one false-positive prediction and one true positive, whereas the old scoring function made three false-positive predictions and one true positive. We repeated this with similar experiments (data available at http://www.fmi.ch/members/andrija.tomovic/exp1.zip) using: MA0052 (two pairs of dependent positions); MA0121 (four pairs of dependent positions); and MA0041 (10 pairs of dependent positions). The accuracy of the new scoring function improved as the number of dependent positions increased. The so-called 'twilight zone' region of the plots also becomes narrower with a smaller density. If there are no dependent positions, then the new and

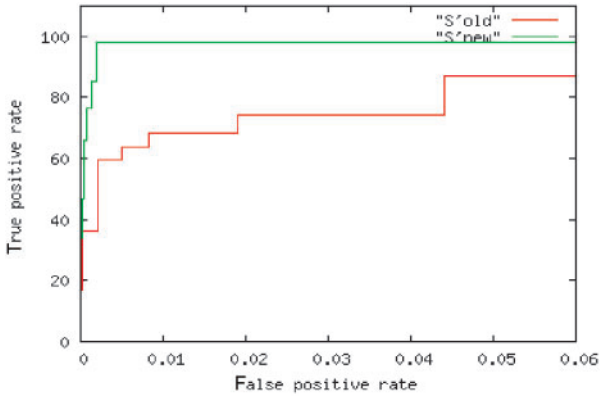


Fig. 2. ROC curves for new and old scoring functions, showing their ability to predict binding sites. The *x*-axis shows the false-positive rate $(FP/(FP+TN)) \times 100$, the *y*-axis shows the true-positive rate $(TP/(TP+FN)) \times 100$.

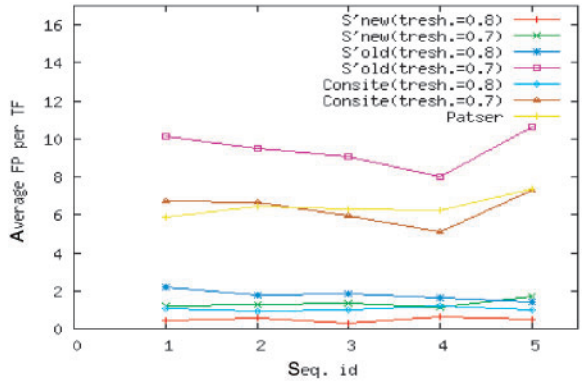


Fig. 3. Average false-positive ratio per TF for different prediction methods.

old scoring functions are the same. We currently only apply our correction for positions that show statistically significant dependencies. If, instead, we factor the observed frequency *P*-scores for all bases, regardless of their significance, then the new function will tend towards the old function because of Equation (12), and the logarithm property ($\log(P(B_i, B_j))$ will tend towards $\log(P(B_i)) + \log(P(B_j))$ but small differences may be observed because of the smoothing parameters. The price for doing this is computational time, and it does not appear to offer any great advantage over the solution we have implemented.

To further evaluate our new scoring function, we generated 1850 random sequences sampled from a third-order Markov model background distribution with lengths from 250 to 500. In 50, we planted binding sites for MA0041 Foxd3, and we then analyzed the true- and false-positive rates for different threshold values using the new and old scoring functions (Fig. 2 and Table Sup8-1 in Supplemental Material 8). Both functions have good scores for true positives, but the new scoring function gave better results. The biggest difference was in the false-positive rate which was much better with the new scoring function. Next, we generated five random sequences sampled from a third-order Markov model background distribution (with lengths from 400 to 600) in which we planted 0–3 binding sites from a set of 15 (all 15 contained dependent positions). The data set is given in Supplemental Material 8. We wanted to measure the accuracy of prediction with the new scoring function and compare it with other available tools and methods (PATSER, ConSite and the old scoring function). Given that almost all of the methods can detect true positives (i.e. they have a high sensitivity), the accuracy of each method should be estimated by its selectivity (false-positive rate). These results are shown in Figure 3 and Table Sup8-2 in Supplemental Material 8. Our new scoring function (22-23) performed best with the smallest number of false positives per nucleotide and per TF. Finally, we analyzed real experimental data. As ConSite had the next best prediction results with the synthetic data, we decided to use it for benchmark comparisons with the experimental data as well. We used a set of genes showing skeletal muscle-specific expression (Wasserman and Fickett, 1998). This set is an updated version from (Defrance and Touzet, 2006) which has been used to evaluate such tools in the past. This dataset includes upstream regions (2000 bp) of nine genes (see Table Sup8-3 in Supplemental Material 8) and six TFs from the JAPSAR database (MA0052, MA0055, MA0056, MA0057, MA0079 and MA0083) which are known to be involved in the regulation of skeletal muscle-specific expression. MA0055's binding sites are not listed in JASPAR, so its detection will be unchanged from the old function (24). We scanned the upstream sequence of the nine genes using all of the TFs from JASPAR. There are 16 TFs (including MA0055) for which there is no binding sequence information, only weight matrices. These will be treated as having independent binding (24), which will have a negative effect on the results from the new scoring function, but is more realistic. However, even with this limitation, the results from the new scoring function are slightly better than those from ConSite (TableSup 8-3 in Supplemental Material 8). The false-positive rate for all nine sequences is smaller with the new scoring function, and the true-positive rate is almost the same. ConSite detected one true positive hit more (for three sequences) than our scoring function with this data set.

4 CONCLUSIONS

In this work, we performed a detailed analysis of dependencies within TF-binding sites. Our conclusion is that we cannot assume that positions are either dependent or independent. This must be tested using one of three proposed statistical tests. Our structural analysis indicates that some of the predicted dependencies agree with 3D structure data from TF–DNA complexes. We propose that the dependencies we have identified should be used in binding-site predictions. Previous attempts at such modeling have required complex tools with many parameters which really require more training data than is currently available. Here, we present a simple way of modeling these dependencies. We demonstrated how to modify existing dependence-free scoring functions to consider dependencies. Such modifications improve prediction quality for TFs

with dependent positions. Our technique does not require complex tools or more training data than scoring functions and models which assume independence. This approach can be used with any scoring function which assumes independence (one such is presented here). We demonstrated this approach using scanning methods for the prediction of TF-binding sites, but it can be applied to work with *ab initio* methods and different methods of prediction which incorporate comparative genomic analysis (phylogenetic footprinting conservation).

ACKNOWLEDGEMENTS

We would like to thank Prof Gill Bejerano, Prof Frank Hampel, Dr Hans-Rudolf Roth, Dr Michael Stadler and Prof Akinori Sarai for useful discussions and advice. This work was supported by the Novartis Research Foundation. Funding to pay the Open Access publication charges was provided by the Novartis Research Foundation FMI.

Conflict of Interest: none declared.

REFERENCES

Agresti,A. (1990) *Categorical Data Analysis*. John Wiley Sons, New York.

Ahmad,S. *et al*. (2006) ReadOut: structure-based calculation of direct and indirect readout energies and specificities for protein-DNA recognition. *Nucleic Acids Res*., **34**, W124–W127.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **2**, 28–36.

Barash,Y. *et al*. (2003) Modeling dependencies in protein-DNA binding sites. In *Proceedings of RECOMB-03*, 28–37.

Bejerano,G. (2003) Efficient exact p-value computation and applications to biosequence analysis. In *Proceedings of RECOMB-03*, 38–47.

Bejerano,G. (2006) Branch and bound computation of exact p-values. *Bioinformatics*, **22**, 2158–2159.

Bejerano,G. *et al*. (2004) Efficient exact p-value computation for small sample, sparse, and surprising categorical data. *J. Comput. Biol*., **11**, 867–886.

Benos,P.V. *et al*. (2002a) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*., **30**, 4442–4451.

Benos,P.V. *et al*. (2002b) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol*., **323**, 701–727.

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res*., **28**, 235–242.

Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol*., **212**, 563–578.

Bulyk,M.L. *et al*. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*., **30**, 1255–1261.

Chiu,D.K. and Kolodziejczak,T. (1991) Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci*., **7**, 347–352.

Cochran,W.G. (1954) Some methods for strengthening the common chi-square tests. *Biometrics*, **10**, 417–451.

Conahan,M.A. (1970) The comparative accuracy of the likelihood ratio and Chi-squared as approximation to the exact multinomial test. Lehigh University, 64.

Davison,A.C. and Hinkley,D.V. (1997) *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.

Day,W.H. and McMorris,F.R. (1992) Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Res*., **20**, 1093–1099.

Defrance,M. and Touzet,H. (2006) Predicting transcription factor binding sites using local over-representation and comparative genomics. *BMC Bioinformatics*, **7**, 396.

Ellrott,K. *et al*. (2002) Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics*, **18** (Suppl. 2), S100–S109.

Gromiha,M.M. (2005) Influence of DNA stiffness in protein-DNA recognition. *J. Biotechnol*., **117**, 137–145.

Gromiha,M.M. *et al*. (2004) Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J. Mol. Biol*., **337**, 285–294.

Gutell,R.R. *et al*. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res*., **20**, 5785–5795.

Hertz,G.Z. *et al*. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci*., **6**, 81–92.

Keich,U. and Nagarajan,N. (2006) A fast and numerically robust method for exact multinomial goodness-of-fit test. *J. Comput. Graph. Stat*., **15**, 779–802

Kel,A.E. *et al*. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*., **31**, 3576–3579.

King,O.D. and Roth,F.P. (2003) A non-parametric model for transcription factor binding sites. *Nucleic Acids Res*., **31**, e116.

Koehler,K.J. (1986) Goodness-of-fit test for log-linear models in sparse contingency tables. *J. Am. Stat. Assoc*., **81**, 483–493.

Koehler,K.J. and Larntz,K. (1980) An empricial investigation of goodness-of-fit statistics for sparse multinomials. *J. Am. Stat. Assoc*., **75**, 336–344.

Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.

Larntz,K. (1978) Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *J. Am. Stat. Assoc*., **73**, 253–263.

Lawrence,C.E. *et al*. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.

Liu,X. *et al*. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *In proceedings of Pac. Symp. Biocomput*., 127–138.

Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*., **32**, W217–W221.

Loots,G.G. *et al*. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res*., **12**, 832–839.

Lu,X.J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*., **31**, 5108–5121.

Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res*., **29**, 2471–2478.

Manly,B.F.J. (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, London.

Marinescu,V.D. *et al*. (2005) MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics*, **6**, 79.

Minka,T. (2003) Bayesian inference, entropy, and the multinomial distribution. Technical Report (Microsoft research).

Perneger,T.V. (1998) What's wrong with Bonferroni adjustments,. *BMJ*, **316**, 1236–1238.

Sandelin,A. *et al*. (2004a) ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*., **32**, W249–W252.

Sandelin,A. *et al*. (2004b) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*., **32**, D91–D94.

Sarai, A. (2006/07) Personal Communication.

Sarai,A. and Kono,H. (2005) Protein-DNA recognition patterns and predictions. *Ann. Rev. Biophys. Biomol. Struct*., **34**, 379–398.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*., **18**, 6097–6100.

Sinha,S. and Tompa,M. (2003) YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*., **31**, 3586–3588.

Sokal,R.R. and Rohlf,F.J. (2003) *Biometry: The Principle and Practice of Statistics in Biological Research*. W.H. Freeman and Company, New York.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Stormo,G.D. *et al*. (1982) Characterization of translational initiation sites in E. coli. *Nucleic Acids Res*., **10**, 2971–2996.

Tompa,M. *et al*. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol*., **23**, 137–144.

Tsunoda,T. and Takagi,T. (1999) Estimating transcription factor bindability on DNA. *Bioinformatics*, **15**, 622–630.

Udalova,I.A. *et al*. (2002) Quantitative prediction of NF-kappa B DNA-protein interactions. *Proc. Natl. Acad. Sci. USA*, **99**, 8167–8172.

van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res*., **31**, 3593–3596.

Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions \which confer muscle-specific gene expression. *J. Mol. Biol*., **278**, 167–181.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet*., **5**, 276–287.

Williams,D.A. (1976) Improved likelihood ratio tests for complete contingency tables. *Biometrika*, **63**, 33–37.

Wolfe,S.A. *et al*. (1999) Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J. Mol. Biol*., **285**, 1917–1934.

Workman,C.T. and Stormo,G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. In *proceedings of Pac. Symp. Biocomput.,* 467–478.

Zhao,X. *et al*. (2005) Finding short DNA motifs using permuted Markov models. *J. Comput. Biol*., **12**, 894–906.

Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.

2.1 Supplementary material 1

for the paper "Position dependencies in transcription factor binding sites"

A. Tomovic and E. J. Oakeley

Derivation 1. **Derivation of formula (14) for calculating the Bayesian factor BF**

In order to test dependencies between positions in transcription factor binding sites by Bayesian hypothesis testing we have to calculate the Bayes factor BF(H_0; H_1), which is similar to Minka (2003) and Zhou and Liu (2004) except that we use different notation and make a small modification.

$$BF(H_0;H_1)=\frac{P(B_i,B_j\mid H_0)P(H_0)}{P(B_iB_j\mid H_1)P(H_1)} \quad (1)$$

If we assume that $P(H_0) = P(H_1) = 0.5$ then (1) will be

$$BF(H_0;H_1)=\frac{P(B_i,B_j\mid H_0)}{P(B_iB_j\mid H_1)} \quad (2)$$

Under the null hypothesis, we have $P(B_i, B_j) = P(B_i)P(B_j)$, and (2) will be

$$BF(H_0;H_1)=\frac{P(B_i\mid H_0)P(B_j\mid H_1)}{P(B_iB_j\mid H_1)} \quad (3)$$

Then, using the fact that:

$$P(B_i\mid H_0)=\int_{\overline{p}} P(B_i,\overline{p}\mid H_0) \quad (4)$$

where $\overline{p}$ is a vector of [P(a,i), P(c,i), P(g,i), P(t,i)],

a conjugate prior for $\overline{p}$ is the Dirichlet distribution:

$$P(\overline{p}\mid a)\sim Dir(a_a,a_c,a_g,a_t)=\frac{\Gamma(\sum_{b_i} a_{b_i})}{\prod_{b_i}\Gamma(a_{b_i})}\prod_{b_i} P(b_i,i)^{a_{b_i}-1} \quad (5)$$

where P(b,i) > 0 and $\sum P(b_i,i)=1$. Given a Dirichlet prior, the joint distribution of B_i and $\overline{p}$ is:

$$P(B_i, \overline{p} \mid a) = \frac{\Gamma(\sum_{b_i} a_{b_i})}{\prod_{b_i} \Gamma(a_{b_i})} \prod_{b_i} P(b_i, i)^{N(b_i,i)+a_{b_i}-1} \quad (6)$$

and the posterior is:

$$P(\overline{p} \mid B_i, a) \sim Dir(N(b_i, i) + a_{b_i}) \quad (7)$$

and, finally, we can calculate:

$$P(B_i \mid H_0) = \int_{\overline{p}} P(B_i, \overline{p} \mid H_0) = \frac{\Gamma(\sum_{b_i} a_{b_i})}{\Gamma(n + \sum_{b_i} a_{b_i})} \prod_{b_i} \frac{\Gamma(N(b_i, i) + a_{b_i})}{\Gamma(a_{b_i})} \quad (8)$$

and likewise for $P(B_j \mid H_0)$.

Then, we need to calculate $P(B_i B_j \mid H_1)$ and this is:

$$P(B_i B_j \mid H_1) = \int_{\hat{p}} P(B_i B_j, \hat{p} \mid H_1) \quad (9)$$

where $\hat{p}$ is a vector of [P(a,a,i,j), P(a,c,i,j), … ,P(t,t,i,j)]

A conjugate prior for $\hat{p}$ is the Dirichlet distribution:

$$P(\hat{p} \mid a) \sim Dir(a_{aa}, a_{ac}, ..., a_{tt}) = \frac{\Gamma(\sum_{b_i, b_j} a_{b_i b_j})}{\prod_{b_i, b_j} \Gamma(a_{b_i b_j})} \prod_{b_i, b_j} P(b_i, b_j, i)^{a_{b_i b_j}-1} \quad (10)$$

where P(b_i,b_j,i,j) and $\sum_{b_i, b_j} P(b_i, b_j, i, j) = 1$. Given a Dirichlet prior, the joint distribution of $B_i B_j$ and $\hat{p}$ is:

$$P(B_i B_j, \hat{p} \mid a) = \frac{\Gamma(\sum_{b_i, b_j} a_{b_i b_j})}{\prod_{b_i, b_j} \Gamma(a_{b_i b_j})} \prod_{b_i, b_j} P(b_i, b_j, i, j)^{N(b_i, b_j, i, j)+a_{b_i b_j}-1} \quad (11)$$

and the posterior is:

$$P(\hat{p} \mid B_i B_j, a) \sim Dir(N(b_i, b_j, i, j) + a_{b_i b_j}) \quad (12)$$

so we can calculate:

$$P(B_i B_j \mid H_1) = \int_{\hat{p}} P(B_i B_j, \hat{p} \mid H_1) = \frac{\Gamma(\sum_{b_i,b_j} a_{b_i b_j})}{\Gamma(n + \sum_{b_i,b_j} a_{b_i b_j})} \prod_{b_i,b_j} \frac{\Gamma(N(b_i,b_j,i,j) + a_{b_i b_j})}{\Gamma(a_{b_i b_j})} \qquad (13)$$

and thus BF can be calculated:

$$BF(H_0;H_1) = \frac{\Gamma(\sum_b a_b)}{\Gamma(n + \sum_b a_b)} (\prod_b \frac{\Gamma(N(b,i) + a_b)}{\Gamma(a_b)}) (\frac{\Gamma(\sum_b a_b)}{\Gamma(n + \sum_b a_b)}) *$$
$$*(\prod_b \frac{\Gamma(N(b,j) + a_b)}{\Gamma(a_b)}) / (\prod_{b_i,b_j} \frac{\Gamma(N(b_i,b_j,i,j) + a_{b_i b_j})}{\Gamma(a_{b_i b_j})} \frac{\Gamma(\sum_{b_i,b_j} a_{b_i b_j})}{\Gamma(n + \sum_{b_i,b_j} a_{b_i b_j})}) \qquad (14)$$

Because we choose $a_{b_i} = \sum_{b_j} a_{b_i b_j}$, we have:

$$BF(H_0;H_1) = \frac{\Gamma(\sum_{b_i,b_j} a_{b_i b_j})}{\Gamma(n + \sum_{b_i,b_j} a_{b_i b_j})} \prod_{b_i} \frac{\Gamma(N(b_i,i) + a_{b_i})}{\Gamma(a_{b_i})} \prod_{b_j} \frac{\Gamma(N(b_j,j) + a_{b_j})}{\Gamma(a_{b_j})} \prod_{b_i,b_j} \frac{\Gamma(a_{b_i b_j})}{\Gamma(N(b_i,b_j,i,j) + a_{b_i b_j})}$$

(15)

The calculation should include only bases b_i, b_j for which $N(b_i,i) \neq 0$ and $N(b_j,j) \neq 0$.

Derivation 2. **Derivation of formula (15): the relationship between BF and mutual information**

It is possible to show (like in Minka, 2003) that there is a relationship between the Bayes factor BF and mutual information M_{ij} if we choose a uniform prior, i.e. $a_k = 1$

Using the fact that $\Gamma(1) = 1$, and the approximation

$$\frac{\Gamma(k)}{\Gamma(n+k)} \approx \frac{\Gamma(k)}{\Gamma(n+1)n^{k-1}} \approx \frac{1}{\Gamma(n+1)}$$

and Stirling's approximation that $\log\Gamma(x+1) \approx x\log x - x$, we get:

$$\log_2(BF(H_0;H_1)) \approx n\sum_{b_i}\frac{N(b_i,i)}{n}\log_2\frac{N(b_i,i)}{n}+n\sum_{b_j}\frac{N(b_j,j)}{n}\log_2\frac{N(b_j,j)}{n}-n\sum_{b_i,b_j}\frac{N(b_i,b_j,i,j)}{n}\log_2\frac{N(b_i,b_j,i,j}{n}$$

$$=-n\sum_{b_i,b_j}\frac{N(b_i,b_j,i,j)}{n}\log_2\frac{N(b_i,b_j,i,j)}{n}\frac{n}{N(b_i,i)}\frac{n}{N(b_j,j)}=$$

$$=-n\sum_{b_i,b_j}P(b_i,b_j,i,j)\log_2\frac{P(b_i,b_j,i,j)}{P(b_i,i)P(b_j,j)}=$$

$$=-nM_{ij}$$

Mutual information and the Bayes factor become more closely related as the sample size n gets higher (because of the approximation of Stirling's formula).

Analytic formula for calculating the hypothetical minimum and maximum for S_{old} and S_{new}

S_{old}^{min} and S_{old}^{max} are hypothetically the minimum and maximum for S_{old}, and S_{new}^{min} and S_{new}^{max} are hypothetically the minimum and maximum for S_{new} calculated by :

$$S_{old}^{min} = \sum_{i=1}^{k} \min_b W_{b,i}$$

$$S_{old}^{max} = \sum_{i=1}^{k} \max_b W_{b,i}$$

$$S_{new}^{min} = \sum_{i=1}^{k_1} \min_b W_{b,i} + \sum_{i=1}^{k_2} \min_{b_1,b_2} W_{b_1,b_2,j_i,j_{i+1}} + \ldots + \sum_{i=1}^{k_m} \min_{b_1,\ldots b_{i+m-1}} W_{b_1,\ldots b_{i+m-1},j_i,\ldots,j_{i+m-1}}$$

$$S_{new}^{max} = \sum_{i=1}^{k_1} \max_b W_{b,i} + \sum_{i=1}^{k_2} \max_{b_1,b_2} W_{b_1,b_2,j_i,j_{i+1}} + \ldots + \sum_{i=1}^{k_m} \max_{b_1,\ldots b_{i+m-1}} W_{b_1,\ldots b_{i+m-1},j_i,\ldots,j_{i+m-1}}$$

REFERENCES

Minka, T. (2003) Bayesian inference, entropy, and the multinomial distribution. Technical report (Microsoft research).

Zhou, Q. and Liu, J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions, *Bioinformatics*, 20, 909-916.

Supplementary material 2

for the paper "Position dependencies in transcription factor binding sites"

A. Tomovic and E. J. Oakeley

Detailed list of statistical values which describe dependence between all positions for each transcription factor from the JASPAR database.

http://promoterplot.fmi.ch/cgi-bin/SupMat/2.xls

Supplementary material 3

for the paper "Position dependencies in transcription factor binding sites"

A. Tomovic and E. J. Oakeley

Detailed list of distributions of different order dependencies.

http://promoterplot.fmi.ch/cgi-bin/SupMat/3.xls

Supplementary material 4

for the paper "Position dependencies in transcription factor binding sites"

A. Tomovic and E. J. Oakeley

Distributions of dependencies according to the structural classification of transcription factors.

http://promoterplot.fmi.ch/cgi-bin/SupMat/4.xls

Supplementary material 5

for the paper "Position dependencies in transcription factor binding sites"

A. Tomovic and E. J. Oakeley

Table Sup5-1. List of PDB and corresponding JASPAR IDs used in the structural analysis. This is actually the intersection of the PDB and JASPAR databases. There are four additional, unused, structures (1IFI, 6PAX, 1FOS and 1H89) which correspond to the JASPAR IDs MA0050, MA0068, MA0099 and MA0100. They were not used because only position weight matrices, and not binding sites, are provided in JASPAR (September, 2006).

PDB ID	JASPAR ID	PDB ID	JASPAR ID
1EGW	MA0001	1YNW	MA0074
1H9D	MA0002	1FJL	MA0075
1AN4	MA0004	1BC8	MA0076
1EGW	MA0005	1PUE	MA0080
1R4I	MA0007	1K60	MA0081
1XBR	MA0009	1J46	MA0084
1K78	MA0014	1SKN	MA0089
1BY4	MA0017	1AN4	MA0093
1NWQ	MA0019	1B8I	MA0094
3HDD	MA0027	1UBD	MA0095
1DUX	MA0028	1GJI	MA0101
2HDC	MA0031	1A1G	MA0103
1AN2	MA0058	1NWQ	MA0102
1AWC	MA0062	1AN2	MA0104
1DSZ	MA0065	1SVC	MA0105
1B72	MA0070	1TSR	MA0106

Table Sup5-2. Structural analysis of co-crystal structures of TFs with DNA (**I** - at least one amino acid that interacted via hydrogen bonds or salt bridges with two binding site positions, both of which were found to be dependent; **II** - at least one amino acid that interacted via hydrogen bonds or salt bridges with two independent binding site positions; **III** - at least one pair of dependent binding site positions with no apparent contact with the transcription factor; **IV** - at least one position of any pair of dependent positions had contact with transcription factors via hydrogen bonds or salt bridges).

Structural characteristic	PDB ID-JASPAR ID
I	1EGW-MA0001, 1AN2-MA0058, 1FJL-MA0075, 1K60-MA0081, 1AN2-MA0104, 1TSR-MA0106
II	1H9D-MA0002, 1AN4-MA0004, 1EGW-MA0005, 1R4I-MA0007, 1BY4-MA0017, 1NWQ-MA0019, 3HDD-MA0027, 2HDC-MA0031, 1AWC-MA0062, 1DSZ-MA0065, 1B72-MA0070, 1FJL-MA0075, 1BC8-MA0076, 1K60-MA0081, 1J46-MA0084, 1AN4-MA0093, 1B8I-MA0094, 1GJI-MA0101, 1NWQ-MA0102, 1SVC-MA0105, 1TSR-MA0106
III	1EGW-MA0001, 1H9D-MA0002, 1EGW-MA0005, 1R4I-MA0007, 1XBR-MA0009, 1K78-MA0014, 1BY4-MA0017, 1NWQ-MA0019, 2HDC-MA0031, 1DSZ-MA0065, 1YNW-MA0074, 1K6O-MA0081, 1NWQ-MA0102, 1TSR-MA0106
IV	1EGW-MA0001, 1H9D-MA0002, 1EGW-MA0005, 1R4I-MA0007, 1K78-MA014, 1NWQ-MA0019, 1DUX-MA0028, 1AN2-MA0052, 1DSZ-MA0065, 1B72-MA0070, 1FJL-MA0085, 1PUE-MA0080, 1J4G-MA0084, 1K6O-MA0081, 1GJI-MA0101, 1NWQ-MA0102, 1A1G-MA0103, 1SVC-MA0105, 1TSR-MA0106

We found six pairs of PDB ID-JASPAR ID in which there was at least one amino acid that interacted via hydrogen bonds or salt bridges with two dependent binding positions. At least one amino acid interacted with two independent binding positions in 21 pairs. However, as discussed before, we anticipate that many of these may become dependent as additional binding site information becomes available. We found 14 pairs in which there was at least one pair of dependent binding site positions with no apparent contact with TFs. There are 19 pairs in which at least one position of any pair of dependent positions had contact with TFs via hydrogen bonds or salt bridges.

Supplementary material 6

for the paper "Position dependencies in transcription factor binding sites"

A. Tomovic and E. J. Oakeley

Distributions of energy Z-scores for direct and indirect readouts.

http://promoterplot.fmi.ch/cgi-bin/SupMat/6.xls

Supplementary material 7

for the paper "Position dependencies in transcription factor binding sites"

A. Tomovic and E. J. Oakeley

Distributions of stiffness for each transcription factor.

http://promoterplot.fmi.ch/cgi-bin/SupMat/7.xls

Supplementary material 8

for the paper "Position dependencies in transcription factor binding sites"

A. Tomovic and E. J. Oakeley

Table Sup8-1. The number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) results for the new and old scoring functions using different threshold values.

THRESHOLD	**new scoring function S'_{new}**				**old scoring function S'_{old}**			
	TP	TN	FP	FN	TP	TN	FP	FN
0.7	47	669,315	138	0	47	640,730	28,723	0
0.8	46	669,362	91	1	46	666,112	3,341	1
0.9	46	669,435	18	1	32	669,326	127	15

For this experiment, we used the dataset available at:

www.fmi.ch/members/andrija.tomovic/expTPTN-supp8.zip and matrixID: MA0041

Table Sup8-2. Average false positive rate per nucleotide and per TF for different methods.

METHOD	**Average FP per nucleotide**	**Average FP per TF**
S'_{new} (threshold = 0.7)	0.04	1.32
S'_{new} (threshold = 0.8)	0.013	0.466
S'_{old} (threshold = 0.7)	0.27	9.453
S'_{old} (threshold = 0.8)	0.05	1.733
ConSite (threshold = 0.7)	0.029	6.33
ConSite (threshold = 0.8)	0.03	1.04
PATSER	0.184	6.453

For this experiment, we used the dataset available at: www.fmi.ch/members/andrija.tomovic/exp-supp8.zip and matrixIDs: MA0006, MA0041, MA0048, MA0052, MA0054, MA0065, MA0066, MA0083, MA0086, MA0091, MA0097, MA0114, MA0116, MA0121 and MA0123

Table Sup8-3. Comparison of the predictions from the new scoring function and ConSite with experimentally verified data.

Gene RefSq ID	**New scoring function S'_{new}**		**ConSite**	
	TP	FP	TP	FP
NM_184041	MA0052, MA0055, MA0056, MA0057, MA0079	78	MA0052, MA0055, MA0056, MA0057, MA0079	81
NM_001927	MA0052, MA0055, MA0056, MA0057, MA0079	74	MA0052, MA0055, MA0056, MA0057, MA0079	80
NM_002479	MA0052, MA0056, MA0055, MA0057, MA0079	84	MA0052, MA0056, MA0055, MA0057, MA0079	87
NM_079422	MA0052, MA0055, MA0056, MA0057, MA0079	79	MA0052, MA0055, MA0056, MA0057, MA0079, MA0083	86
NM_003281	MA0052, MA0055, MA0056, MA0057, MA0079	78	MA0052, MA0055, MA0056, MA0057, MA0079	81
NM_000257	MA0055, MA0056, MA0057, MA0079	75	MA0052, MA0055, MA0056, MA0057, MA0079	77
NM_002471	MA0052, MA0055, MA0056, MA0057, MA0079	78	MA0052, MA0055, MA0056, MA0057, MA0079	82
NM_001100	MA0052, MA0055, MA0056, MA0057, MA0079	77	MA0052, MA0055, MA0056, MA0057, MA0079, MA0083	80
NM_005159	MA0052, MA0055, MA0056, MA0057, MA0079	77	MA0052, MA0055, MA0056, MA0057, MA0079	84

Note: MA0055 is not available in our database, because there is no sequence data in the JASPAR database (September, 2006), only profile. However, searching using profile it is possible to find it.

Supplementary material 9

Web-based implementation of the new scoring function (Figure 1), publicly available from http://promoterplot.fmi.ch/cgi-bin/dep.html.

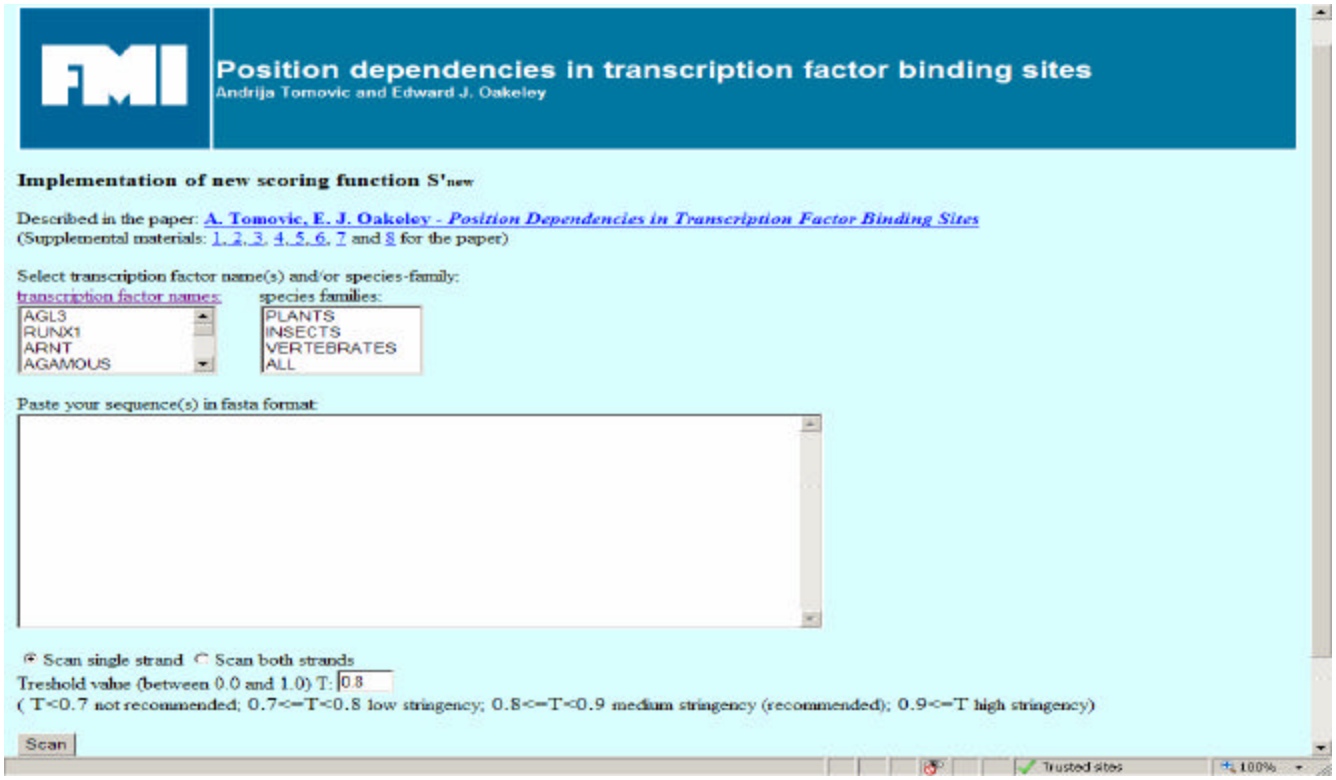


Figure 1. *Web-based tool for the computational prediction of transcription factor binding sites*

3. Computational structure analysis: multiple proteins bound to DNA (paper III)

Computational structural analysis of protein-protein and single protein-DNA interactions (binary complexes) is well documented. The increased number of structures with multiple proteins bound to DNA gives a good opportunity for performing a descriptive data-mining study on that kind of structure in order to extract useful information (structural, physical-chemical and thermodynamic parameters). This information can help further the theoretical understanding of DNA-protein interactions and, in addition, can be helpful for generating hypothetical complexes, predicting DNA-binding specificities, designing novel DNA-binding proteins and predicting the assembly of transcription factor complexes.

This chapter contains a paper which is currently under review (submitted). All supplementary materials (additional files) for this paper are available in this chapter, and the implementation of the algorithm presented is publicly available from http://promoterplot.fmi.ch/Collision1/

OPEN ACCESS Freely available online

PLoS one

Computational Structural Analysis: Multiple Proteins Bound to DNA

Andrija Tomovic*, Edward J. Oakeley

Friedrich Miescher Institute for Biomedical Research, Novartis Research Foundation, Basel, Switzerland

Abstract

Background: With increasing numbers of crystal structures of protein:DNA and protein:protein:DNA complexes publically available, it is now possible to extract sufficient structural, physical-chemical and thermodynamic parameters to make general observations and predictions about their interactions. In particular, the properties of macromolecular assemblies of multiple proteins bound to DNA have not previously been investigated in detail.

Methodology/Principal Findings: We have performed computational structural analyses on macromolecular assemblies of multiple proteins bound to DNA using a variety of different computational tools: PISA; PROMOTIF; X3DNA; ReadOut; DDNA and DCOMPLEX. Additionally, we have developed and employed an algorithm for approximate collision detection and overlapping volume estimation of two macromolecules. An implementation of this algorithm is available at http://promoterplot.fmi.ch/Collision1/. The results obtained are compared with structural, physical-chemical and thermodynamic parameters from protein:protein and single protein:DNA complexes. Many of interface properties of multiple protein:DNA complexes were found to be very similar to those observed in binary protein:DNA and protein:protein complexes. However, the conformational change of the DNA upon protein binding is significantly higher when multiple proteins bind to it than is observed when single proteins bind. The water mediated contacts are less important (found in less quantity) between the interfaces of components in ternary (protein:protein:DNA) complexes than in those of binary complexes (protein:protein and protein:DNA).The thermodynamic stability of ternary complexes is also higher than in the binary interactions. Greater specificity and affinity of multiple proteins binding to DNA in comparison with binary protein-DNA interactions were observed. However, protein-protein binding affinities are stronger in complexes without the presence of DNA.

Conclusions/Significance: Our results indicate that the interface properties: interface area; number of interface residues/atoms and hydrogen bonds; and the distribution of interface residues, hydrogen bonds, van der Walls contacts and secondary structure motifs are independent of whether or not a protein is in a binary or ternary complex with DNA. However, changes in the shape of the DNA reduce the off-rate of the proteins which greatly enhances the stability and specificity of ternary complexes compared to binary ones.

Citation: Tomovic A, Oakeley EJ (2008) Computational Structural Analysis: Multiple Proteins Bound to DNA. PLoS ONE 3(9): e3243. doi:10.1371/journal.pone.0003243

Editor: Mark Isalan, Center for Genomic Regulation, Spain

Received July 8, 2008; **Accepted** August 24, 2008; **Published** September 19, 2008

Funding: This work was supported by the Novartis Research Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: andrija.tomovic@fmi.ch

Introduction

DNA-binding proteins are important for the regulation of many crucial cellular processes (including transcription, recombination, and replication). The number of DNA-binding proteins known is very small compared to the number of regulatory controls they must provide within the nucleus. The problem is solved, at least in part, by the construction of higher-order regulatory complexes composed of multiple proteins. Structural analyses of such complexes may enable us to model the forces driving their assembly and stability which in turn may help us to understand these processes better. Such an understanding may help in predicting DNA-binding specificities. Transcription factors, a large subclass of DNA-binding proteins, are known to act cooperatively in the regulation of gene expression [1–7]. Their complexes can include both DNA and non-DNA-binding factors. The DNA-binding factors may be located either remotely (at some distance) or adjacent (with direct contacts) to their promoters [5].

Thanks to a large number of recent X-ray and NMR structures of protein:protein, protein:DNA, and protein:RNA complexes, a lot of valuable information about the general features of such complexes has been discovered [8–23]. These results indicate that it is very difficult to find universally characteristic rules which can describe all protein-protein, protein-DNA, and protein-RNA interactions. However, some general principles have been deduced. For example, Lys or Arg pair preferentially with any nucleotide in both protein:DNA and protein:RNA complexes [16]; two-thirds of all protein-DNA interactions involve van der Waals contacts, compared to about one-sixth involving hydrogen bonds [18]; on average protein-protein interface has approximately the same non-polar character as the protein surface as a whole and carries somewhat fewer charged groups (however, some

interfaces are significantly more polar and others more non-polar than the average) [17].

The current work comprises a structural analysis of macromolecular assemblies where several proteins are bound to DNA, using data from the Protein Data Bank (PDB) [24]. We analyzed the following chemical and physical properties: the size of interfaces between any two components; the number of residues/atoms involved in contacts between components; residue interface propensities and chemical composition; water-mediated contacts in interfaces; secondary structure motifs in interfaces; and interactions between amino acid side chains either with the DNA or with another protein in the complex. Some of these interface properties for ternary/quaternary complexes (i.e. complexes involving two/three proteins bound to DNA) have been compared with those obtained from binary complexes. One possible hypothesis why the above-mentioned protein-DNA and protein-protein interface properties are expected to depend on the number of proteins in a complex is that when two proteins are free (not bound to DNA) they are more able to find the best patches (on both proteins) to produce the most stable complexes possible, with the highest affinity between components. However, when one protein is bound to DNA then there is a spatial limitation in the movements that are possible in order to find the best interface patches (on both proteins) in order to make stable complexes. This is one possible explanation why protein-protein interface properties can be expected to be different in protein:protein and in protein:protein:DNA complexes. A possible implication is that (if properties are similar or the same) actually two DNA-binding proteins bind first to each other and then bind to DNA together (as a complex). A similar hypothesis can be derived for protein-DNA interfaces in protein:DNA and in protein:{protein+}:DNA complexes. One might suppose that these interfaces can be different, because when one protein binds to DNA there is a higher degree of freedom (rotational, translational) than when one protein should bind to a previously-made protein:DNA complex. This is useful (from a theoretical point of view) for better understanding protein-DNA interactions which frequently involve complexes of multiple proteins. In addition, this can be useful (from a practical point of view) for the possible modelling of such complexes (their prediction, prediction of order of processes, modelling cis-regulatory modules, etc). In addition the nature of protein-protein interface and protein-DNA interface might be different that there is no any competition between them. This aspect can be also considered with this kind of analysis performed in this paper. In this work we have also calculated and compared, the conformational change of DNA in binary complexes (i.e. single protein-DNA complexes) and ternary/quaternary complexes (protein-protein-DNA/protein-protein-protein-DNA). Next, we analyzed protein-protein and protein-DNA energy binding affinity in protein-protein, single protein-DNA and multiple proteins-DNA complexes using several different tools. In addition, we analyzed and compared the thermodynamic stabilities of these complexes. We have provided an algorithm, and its web-based implementation, for calculating overlapping interface volumes and the number of interface atoms in collision between any two components (macromolecules) from a 3D complex stored in a pdb file.

Results and Discussion

We have performed computational structural analysis and present herewith some general features we have observed about macromolecular assemblies of multiple proteins bound to DNA. The following tools were used in our analysis: PISA [25,26]; PROMOTIF [27]; X3DNA [28]; ReadOut [29]; DDNA [30] and DCOMPLEX [31]. Additionally, we have developed and used an algorithm for collision detection and overlapping volume of two macromolecules. Web-base implementation of the algorithm is freely available from http://promoterplot.fmi.ch/Collision1/ (see Materials and Methods for details). All data sets, used in this study, are from the PDB database (see Materials and Methods for a definition of data sets used in this study).

Physical properties of interfaces

Do physical properties of interfaces depend on the number of units in macromolecular assemblies? Are there any differences in physical properties of interfaces among protein:protein:DNA, protein:DNA and protein:protein complexes? In order to answer these questions, we performed analysis of physical interface properties of different macromolecular assemblies.

The number of interfaces in the dataset MutliProteins:DNA together with their structural characteristics is summarized in Table 1.

A detailed list of 52 protein-protein and 87 protein-DNA interfaces is given in Table S1. These values represent the sample sizes for the following hypothesis tests between protein-protein and protein-DNA interactions: There was no significant difference in average interface surface sizes (student's t-test, p-value = 0.69); nor the average number of interface residues (student's t-test, p-value = 0.76) nor the average number of atoms (p-value = 0.41). Based on this we can conclude that protein-protein and protein-DNA interfaces have similar average sizes and numbers of residues/atoms involved in their interactions in protein:protein:DNA complexes. La Conte et al. [17] found that most protein-protein interface areas are in the range of 1200–2000 $Å^2$. They consider the total area on both components (without dividing by 2 to make the average area) as shown in formula (2). The protein-protein and protein-DNA interface areas for protein:protein:DNA complexes are also to this range (Table 1). The average area of protein-protein interfaces of complexes in the group-MultiProteins:DNA and the average area of protein-protein interfaces of complexes in the group-Protein:Protein we observe

Table 1. Descriptive statistics of interfaces.

Interface type	Number of interfaces	Average size of interface ($Å^2$)±SE	Average number of interface residues*±SE	Average number of interface atoms*±SE	Average number of intermolecular H-bonds±SE	Average number of intermolecular salt bridges±SE
Protein-protein	52	929.84±179.4	49.5±8.4	190.9±36.0	9.36±3.7	4.08±0.7
DNA-protein	87	1002.3±56.5	52.2±2.9	222.2±12.5	18.0±1.1	0.0±0.0

Descriptive statistics of protein-protein and protein-DNA interfaces of complexes from group-MultiProteins:DNA.
*For both components together in interface.
doi:10.1371/journal.pone.0003243.t001

was comparable to those reported by Chakrabarti and Janin [9]. The DNA interface area sizes reported in Table 1 are comparable with those reported in studies considering only single protein-DNA complexes [15,21]. The number of residues/atoms in protein-protein interfaces in this study was also comparable to previous studies [9,17]. The situation is similar if we compare protein-DNA interfaces of protein:protein:DNA complexes with protein-DNA interfaces of protein:DNA complexes [15,21].

Based on this we can conclude that average interface size and the average number of interfaces residues/atoms between two macromolecules (DNA, protein) in any kind of complex (protein:-protein, protein:DNA, protein:protein:DNA) are approximately the same. In addition, it appears that these physical properties are not influenced by the number of subunits in the complex.

Distribution of hydrogen bonds in interfaces

The purpose of this section was to investigate differences in distributions of hydrogen bonds between interfaces of macromolecular assemblies. There is a statistically significant difference in the average number of intermolecular hydrogen bonds (H-bonds) between protein-protein and DNA-protein interfaces (student's t-test, p-value$<$0.0001). The number of H-bonds observed in previous protein-protein studies (mean 10.1 ± 0.5) [17] is comparable to those reported in this study for group-MultiProteins:DNA (Table 1). The situation is similar if we compare protein-protein-DNA verses protein-DNA interfaces [15,21]. The small observed variations are due to small variations in the interface areas as the number of hydrogen bonds is dependent on this area.

In Table S2 we report the numbers of hydrogen bonds observed between the 20 amino acids and the four bases or the backbone of the DNA for the complexes listed in the group-MutliProteins:DNA. We found that H-bond pairs were significantly different from random (Fisher's test, $p<10^{-6}$). The most favoured amino acid-DNA base H-bond is ARG-G. In Figure S1 we report the distribution of H-bonds between the DNA bases and the bound proteins in group-MutliProteins:DNA. 65.69% of all H-bonds where between protein side chains and the DNA backbone (Figure S1). Those H-bonds are not expected to confer specificity of binding but rather assist in complex stability. Most amino acids involved in H-bonds between the proteins and DNA (complex from group-MultiProteins:DNA) are positively charged, presumably because of the negative charge of DNA (Figure S2). For the H-bonds at the protein-protein interfaces, the situation is different: negative and positively charged amino acids have an approximately equal frequency due to the need to pair charges in electrostatic interactions between donator and acceptor sites in the two proteins. Very similar distributions of H-bonds are found in groups –SingleSameProtein:DNA and –SubSetMultiProteins:DNA (Table S3, Table S4, Figure S3, Figure S4).

Most H-bonds (53.3%) are made with phosphate groups of the DNA at the protein:DNA interfaces. Very few H-bonds (12%) are made with deoxyribose (Figure S1). This situation is the same as that reported by Lejeune et al. [16] and Luscombe et al. [18] for protein-DNA interactions. The distribution of H-bonds between the participating amino acids and the DNA is given in Table S2. Entries in Table S2 that diverge from the expected distribution (favoured amino acid-base H-bonds) are also similar to those observed by Luscombe et al. [18].

Distributions of interface residues

In this section we present results about distributions of interface residues. We investigate if distributions of interface residues dependent on the number of units in the complex and if there are any differences in residue distributions between binary and ternary complexes (protein:protein:DNA, protein:DNA, protein:-protein). The amino-acid propensities for the protein-protein and protein-DNA interfaces for complexes from the group-Multi-Proteins:DNA are shown in Figure S5. For protein-DNA interfaces, ARG and LYS have the highest propensity values (>1.2), which indicates that they occur greater than 20% higher frequently in the interfaces than in the whole dataset. On other hand, many amino acids (ALA, ASP, CYS, GLN, GLU, ILE, LEU, MET, PHE, PRO, and VAL) are disfavoured in the interactions sites. For protein-protein interfaces, the situation is different and MET is the most favoured residue at interaction sites. In Figure S6 we report the distribution of amino acids involved in protein-protein and protein-DNA interfaces in the complexes from the group-MultiProteins:DNA. Aliphatic amino acids are dominant in protein-protein interactions, while positively charged amino acids are the most involved in protein-DNA interactions. Those two distributions are significantly different, with a p-value$<$0.0001 (Chi-square multinomial test). The complexes in group-MutliProteins:DNA have a number of van der Waals interactions between the amino acids in the proteins and either the DNA bases or backbone that is significantly different from random (Table S5, Fisher's p-value$<5\times10^{-6}$). In order to determine which of the pairings are different from expected, we performed individual Fisher's tests on each pair. The distributions of interface residues for protein-DNA interfaces of the complexes in the groups-SubSetMultiProteins:DNA and –SingleSameProtein:DNA are reported in Table S6 and Table S7.

Protein-protein interfaces are more hydrophobic than protein-DNA interfaces (they contain significantly more aliphatic amino acids, see Figure S6 for details). Protein-protein interfaces have many more negatively charged amino acids and far fewer positively charged amino acids than protein-DNA interfaces. All these interface parameters give an indication of the overall polar nature of protein-DNA interfaces. Given that the DNA molecule surface is negatively charged, it is perhaps not surprising that it favours positively charged protein surface patches.

The frequency distributions of amino acids in protein-DNA interaction sites in this study from the group-MultiProteins:DNA are similar to those reported by Lejeune [16] (Figure S5 and Figure S6).

Distribution of interface structural motifs

We investigated if the distributions of structural motifs in interfaces of components in ternary (protein:protein:DNA) complexes are different from those in binary complexes (protein:protein and protein:DNA). In order to answer on this question we calculate the propensity values for protein-protein and protein-DNA secondary structure motifs from the group-MultiProteins:DNA (shown in Figure 1). The most favoured protein-DNA interface motif in is the helix, and the least favoured motifs are γ-turns, β-strands, and β-hairpins. At protein-protein interfaces, the least favoured secondary structure motif is the β-bulge. The distributions of secondary structure motifs between protein-protein and protein-DNA interfaces are significant different (Chi-square multinomial goodness-of-fit test, p-value$<$0.01). For protein-DNA interfaces, the dominant structural motif is the helix. This result is consistent with the observation that many DNA binding sites on proteins are comprised of helix motifs [32]. The distribution of secondary structure motifs in protein-protein interfaces for the complexes used in this study (group-MultiProteins:DNA, Figure 1) is similar to that observed by Guharoy and Chakrabarti [33] who observed that the contribution of β-strands is lower than that of helixes and that non-regular structural motifs appear in large numbers.

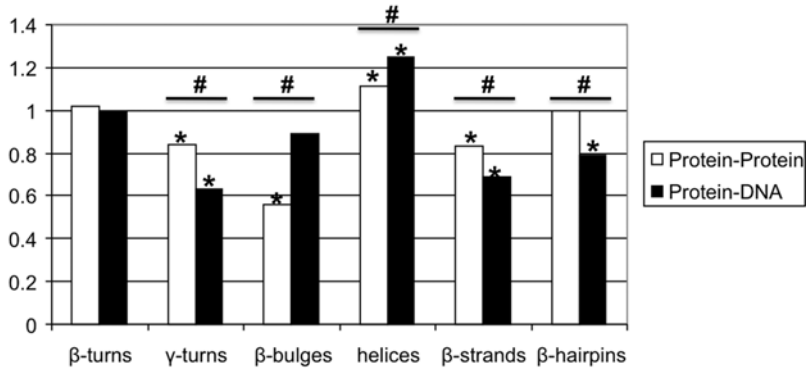


Figure 1. Secondary structure motif propensities. Secondary structure motif propensities for protein-protein and protein-DNA interfaces. Propensity values which are significantly different from 1 (either above or below), evaluated by the statistical bootstrapping method, are marked with "*". Significant statistical differences between motif propensities of protein-protein and protein-DNA interfaces are marked with "#".
doi:10.1371/journal.pone.0003243.g001

All previous results (from this and previous subsections) can be summarized in the form:

$$\begin{aligned} &X_{\text{protein}-\text{protein}}(\text{protein} : \text{protein}) \\ &+X_{\text{protein}-\text{DNA}}(\text{protein} : \text{DNA}) \\ &\approx X_{\text{protein}-\text{protein}}(\text{protein} : \text{protein} : \text{DNA}) \\ &+X_{\text{protein}-\text{DNA}}(\text{protein} : \text{protein} : \text{DNA}) \end{aligned} \quad (1)$$

where $X_{\text{protein-protein}}$ (C) and $X_{\text{protein-DNA}}$ (C) represent one of the following interface parameters: area, number of residues, number of atoms, number of H-bonds, distribution of residues, distribution of H-bond partners or the distribution of structural interface motifs in either protein-protein or protein-DNA interfaces respectively where complex C is either a protein:protein, a protein:DNA or a protein:protein:DNA complex. Formula (1) can be easily be expanded to cover quaternary complexes (protein:protein:protein:DNA) as well, but for clarity we have only represented the case for ternary complexes.

It is apparent from formula (1) that interface parameters under discussion, for complexes composed of multiple proteins bound to DNA, can be estimated from protein-protein and single protein-DNA complexes alone. A more precise variant of formula (1), for example in the form of a regression equation, would be possible to derive if we had crystal structures of the same protein in all three states: protein:protein; protein:DNA and protein:protein:DNA.

Our results indicate that the physical properties of protein:protein and protein:DNA complexes, such as interface area, number of interface residues/atoms and hydrogen bonds and the distribution of interface residues and secondary structure motifs are no different in binary or ternary complexes. Thus, if we have two (or more) proteins which bind together, there will be no influence on these interface parameters of their DNA-binding interface when they bind together as a complex to DNA. This claim is not related to the energy of these interactions and it is expected that the interaction rate constants will not be the same for binary and multiple proteins complexes. If two DNA binding proteins can also bind to each other then this will tether them in the vicinity of the DNA such that when one of the proteins binds to DNA the second will have a faster on-rate because it will have a shorter distance to diffuse to find its binding site thus maintain a higher effective local concentration around the DNA. A detailed analysis of rate constants cannot unfortunately be made from crystal structures which are by definition static snapshots of this dynamic process.

Water molecules in protein-protein and protein-DNA interactions

It has been discussed that water content and water mediated contacts in the protein-DNA interface are important components of protein-DNA interactions [34,35]. Protein-protein and protein-DNA interfaces contain significant quantities of water [36]. Structural and biochemical data indicate that water-mediated interactions are important for the stability and specificity of recognition, despite the fact that interface solvent molecules exchange rapidly with the bulk solvent [36]. We wanted to evaluate the differences between water mediated contacts at protein-DNA interfaces in protein:DNA complexes (single proteins bound to DNA) and in protein:protein:DNA complexes (multiple proteins bound to DNA). The average number of water mediated contacts between the protein-DNA interfaces of protein:protein:DNA complexes is $\sim 11.82 \pm 1.3$ (Table S8). This is markedly different from the value of 28 reported for protein:DNA complexes previously [36]. Similarly, we compared the water mediated contacts in the protein-protein interfaces of protein:protein and protein:protein:DNA complexes. The average number of water molecules for protein-protein interfaces of complexes in the group-MultiProteins:DNA was $\sim 4.9 \pm 0.83$ (Table S8), as compared to ~ 22 for protein-protein interactions in binary protein:protein complexes reported by [36].

These results suggest that water mediated contacts in interfaces of components in protein:protein:DNA complexes play less important role in the stability and specificity of recognition then in interfaces of components in the binary protein:protein and protein:DNA complexes. However, as we discussed later in the text there are other factors which are more important for stability and specificity of component recognition in protein:protein:DNA complexes.

DNA distortion

In order to check if DNA structural deformation is higher when multiple proteins bind to DNA we performed computational structural analysis of DNA structures. DNA distortion was measured by calculating the root-mean-square deviation (rmsd) when each DNA structure was fitted onto its corresponding canonical A-DNA or B-DNA structure. Distributions of rmsd values for all complexes from the groups MultiProteins:DNA (black bars) and SingleSameProtein:DNA (white bars) were calculated (Figure 2). Statistical analysis of these results showed a significant difference in means of rmsd values (student's t-test with

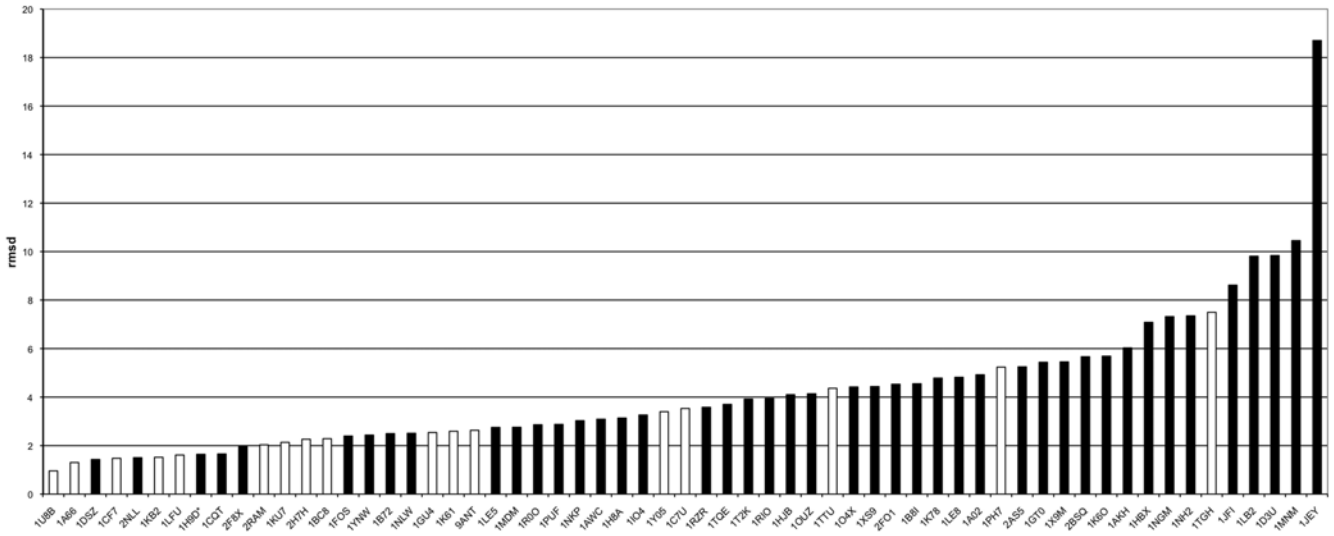


Figure 2. Distribution of rmsd values for measuring DNA distortion. Distribution of rmsd values calculated from fitting each DNA structure in the complexes from group-MultiProteins:DNA (black bars) and group-SingleSameProtein:DNA (white bars) to a corresponding canonical B-DNA.
doi:10.1371/journal.pone.0003243.g002

equal or unequal variance as appropriate, p-value$<$0.02) calculated for all complexes from the groups –MultiProteins:DNA, -SingleProtein:DNA and –SingleSameProtein:DNA calculated after fitting each DNA structure onto the corresponding canonical A-DNA and B-DNA structures (Table 2). Further information for each complex is given inTable S9, S10, S11 and S12. The rmsd values for the group-SubMultiProteins:DNA are the same as those for the group-MultiProteins:DNA.

The rmsd values of the group SubSetMultiProteins:DNA, including comparisons with the group SingleSameProtein:DNA, are given in Table S13. DNA distortion, however, is significantly higher when multiple proteins are bound to the DNA (Figure 2, Table 2, Table S13). It has been reported that when a single protein binds to DNA it results in a higher rmsd (conformational change) than that seen in the unbound DNA structure [15]. Here we reported that there are also further conformational changes to the structure of DNA which are induced when multiple proteins bind to it.

Energetic properties of interfaces

The energetic properties of cooperatives are useful for understanding of how the essential macromolecular machines of cellular function are assembled and how they work [37]. We analyzed energetic and thermodynamic properties of different mulitcomponent complexes (protein:protein:DNA, protein:DNA, protein:protein). In Table 3 we report the free energy of dissociation (ΔG^{diss}) and the free energy of solvation (ΔG^{int}) in kJ/mol for complexes from the four groups –MultiProteins:DNA, -SubMultiProteins:DNA, -SingleProtein:DNA, and –SingleSameProtein:DNA. In Table 4 we also report energy Z-score values for direct and indirect readouts for the three groups –MultiProteins:DNA, -SubMultiProteins:DNA and –SingleProtein:DNA. The p-values in Table 3 were obtained by comparing the means of ΔG^{int}, ΔG^{diss} and the Z-scores for the direct and indirect readouts using the student's t-test (with equal or unequal variance as appropriate). We could not calculate energy Z-scores for the indirect readouts of the group SubMultiProteins:DNA because the DNA structure is the same for each complex, so the calculated Z-scores would also be the same. Detailed lists of the ΔG^{int}, ΔG^{diss} and Z-scores for both the direct and indirect readouts of each complex and each group are available in Table S14, S15, S16, S17, S18, S19, S20, S21, S22 and S23.

Table 4 shows the average protein-DNA energy binding affinity in kJ/mol for the MultiProteins:DNA, SubMultiProteins:DNA, SingleProtein:DNA and SingleSameProtein:DNA groups; the average protein-DNA overlapping volume (in Å^3) and the number of atoms in collision at the protein-DNA interfaces. All values were compared against the MultiProteins:DNA group and a student's t-test was used to calculate the p-values. Further information on these parameters can be found in Table S24, S25, S26, S27 and S28.

The average protein-protein binding energy for complexes from the MultiProteins:DNA group (which are bound to DNA) is significantly smaller (student's t-test, p-value = 0.05) than that of

Table 2. Measuring DNA distortion.

Dataset of complexes	Average rmsd (±SE) from A-DNA	Average rmsd (±SE) from B-DNA
Group-MultiProteins:DNA	8.26±0.4	4.71±0.5
Group-SingleProtein:DNA	5.94±0.2(p$<$0.001)	3.44±0.2 (p = 0.007)$^{\#}$
Group-SingleSameProtein:DNA	6.66±0.6 (p = 0.02)	2.87±0.4 (p = 0.004)$^{\#}$

Average rmsd values calculated from fitting each DNA structure in the complexes from group –MultiProteins:DNA, -SingleProtein:DNA, and –SingleSameProtein:DNA to a corresponding canonical A-DNA and B-DNA.
p-values are calculated in comparison with Group A and obtained using the one-tailed Student's t-test.
$^{\#}$unequal variance.
doi:10.1371/journal.pone.0003243.t002

Table 3. Complex energies.

Dataset of complexes	Average (±SE) solvation energy ΔG^{int} (kJ/mol)	Average (±SE) ΔG^{diss} (kJ/mol)	Average (±SE) energy Z-score for direct readout	Average (±SE)energy Z-score for indirect readout
Group-MultiProteins:DNA	−234.61.03±18.4	50.41±6.0	−2.81±0.2	−2.36±0.1
Group-SubMultiProteins:DNA	−123.21±9.8 (p<0.001)[#]	47.19±4.9 (p = 0.34)	−1.71±0.2 (p<0.001)	—
Group-SingleProtein:DNA	−114.49±8.6 (p<0.001)[#]	48.52±5.3 (p = 0.41)	−1.84±0.3 (p = 0.005)[#]	−2.14±0.1 (p = 0.13)
Group-SingleSameProtein:DNA	−99.79±15.0 (p<0.001)[#]	31.06±6.5 (p = 0.03)	−1.34±0.3 (p<0.001)[#]	−1.48±0.3 (p = 0.007)

Average solvation energy (kJ/mol), free energy barrier of assembly dissociation (kJ/mol), and energy Z-scores for direct and indirect readouts for groups –MultiProteins:DNA, -SubMultiProteins:DNA, -SingleProtein:DNA and –SingleSameProtein:DNA.
p-values are calculated in comparison with Group-MultiProteins:DNA and obtained using the one-tailed Student's t-test.
[#]unequal variance.
doi:10.1371/journal.pone.0003243.t003

Table 4. Affinity of components.

Dataset of complexes	Average (±SE) protein-DNA energy binding affinity (kJ/mol)	Average (±SE) protein-DNA overlapping volume (Å^3)	Average (±SE) number of atoms in collision in protein-DNA interfaces
Group-MultiProteins:DNA	−39.05±0.9	4.26±0.8	32.06±4.1
Group-SubMultiProteins:DNA	−30.93±0.5 (p<0.001)[#]	2.04±0.3 (p = 0.007)[#]	15.44±1.9 (p<0.001)[#]
Group-SingleProtein:DNA	−33.20±0.6 (p<0.001)	3.17±0.56 (p = 0.13)	20.45±1.8 (p = 0.006)[#]
Group-SingleSameProtein:DNA	−32.79±0.9(p<0.001)[#]	2.313±0.8 (p = 0.04)[#]	15.5±3.3 (p = 0.001)[#]

Average protein-DNA energy binding affinity (kJ/mol), interface overlapping volume (Å^3) and average number of interface collision atoms for groups –MultiProteins:DNA, -SubMultiProteins:DNA, -SingleProtein:DNA and –SingleSameProtein:DNA.
p-values are calculated in comparison with Group-MultiProteins:DNA and obtained using the one-tailed Student's t-test.
[#]unequal variance.
doi:10.1371/journal.pone.0003243.t004

complexes from group-Protein:Protein (Table 5). The average solvation energy (ΔG^{int}) and free energy barrier of assembly dissociation (ΔG^{diss}) for protein-protein complexes from group–MultiProteins:DNA is, respectively, smaller and larger (student's t-test, p-value<0.001) than that found for complexes from group-Protein:Protein (Table 5). A list of protein-protein binding affinities for every complex in the MultiProteins:DNA and Protein:Protein groups may be found in Table S29–S30.

The energetic properties of protein-DNA interfaces of the complexes in group-SubSetMultiProteins:DNA, including their comparisons with corresponding values from group-SingleSameProtein:DNA, are given in Tables S31 and S32.

The free energy barrier of assembly dissociation (ΔG^{diss}, Table 3) is higher for complexes involving multiple proteins bound to DNA (MultiProteins:DNA) than those involving only single protein-DNA complexes (SubMultiProteins:DNA, SingleProtein:DNA and SingleSameProtein). The SingleSameProtein:DNA and the SubMultiProteins:DNA groups both contain proteins which are also components of the complexes found in the MultiProteins:DNA group, but the SubMultiProteins:DNA group was formed by manually removing the extra protein units from the complexes of group-MultiProteins:DNA in order to get single protein-DNA complexes. We see that in comparison with the SingleSameProtein:DNA group, complexes in the MultiProteins:DNA group have significantly ($p = 0.03$, student's t-test) higher free energy barriers of assembly dissociation (ΔG^{diss}). This means that multiple proteins-DNA complexes are more thermodynamically stable than single protein-DNA complexes. Comparing the MultiProteins:DNA group to the three other groups (SubMultiProteins:DNA, SingleProtein:DNA, and SingleSame-

Table 5. Protein-protein interfaces energies.

Dataset of complexes	Average (±SE) protein-protein binding free energy (kJ/mol)	Average (±SE) solvation energy ΔG^{int} (kJ/mol)	Average (±SE) ΔG^{diss} (kJ/mol)
Group-MultiProteins:DNA	−56.27±6.3	−234.61.03±18.4[*]	50.41±6.0[*]
Group-Protein:Protein	−67.20±2.3 (p = 0.05)[#]	−81.937±10.1 (p<0.001)[#]	8.22±2.9 (p<0.001)[#]

Average protein-protein binding free energy (kJ/mol), average solvation energy (kJ/mol) and average free energy barrier of assembly dissociation (kJ/mol) for protein-protein complexes from group –MultiProteins:DNA and –Protein:Protein.
p-values are calculated in comparison with Group-MultiProteins:DNA and obtained using the one-tailed Student's t-test.
[#]unequal variance.
[*]calculated for the whole complex (the same values as in Table 3).
doi:10.1371/journal.pone.0003243.t005

Protein:DNA), we find a significantly smaller free energy (student's test, p-value$<$0.001, Table 3) of solvation gain upon complex formation (ΔG^{int}). The same result was found when comparing the MutliProteins:DNA group to the SubSetMultiProteins:DNA group (Table S31).

The energy Z-scores for direct and indirect readouts (conformational energy) have more negative values for complexes with multiple proteins bound to DNA (Table 3 and Table S31). More negative Z-scores mean that the target DNA sequence fits into a given protein structure better [29]. Therefore, DNA-binding proteins fit their targets better when they form a ternary complex with DNA. The Z-score also indicates that ternary complexes may be more stable than binary ones. The binding energy affinity, overlapping volume and number of atoms in collision (Table 4) is significantly higher in protein-protein-DNA complexes than in protein-DNA complexes. Differences in overlapping volume and number of atoms in collision are due not only to the bigger interface area (twice protein:DNA), but also to the higher affinity of multiple proteins binding (interface area sizes for the SingleProteins:DNA, SingleSameProteins:DNA and –SubMultiProteins:DNA groups are similar, butthe SingleProtein:DNA and SingleSameProtein:DNA groups have higher protein-DNA binding affinities, overlapping volumes and numbers of atoms in collision than those in the SubMultiProteins:DNA group, Table 4 and Table S32). Cis-modules that contain transcription factor binding sites (cis-motifs) of transcription factors which make direct physical contact with each other have higher DNA-binding affinities than cis-modules that contain transcription factor binding sites (cis-motifs) of factors without direct mutual contacts. This information may be used for the prediction of cis-regulatory motifs/modules in the following way: if we say that the value of a scoring function for binding sites which are close to one another (where there might be the physical contact between corresponding transcription factors) may have a lower threshold value than a threshold which should be used for scoring function for binding sites that are further away (where there might not be the physical contact between corresponding transcription factors). Modelling DNA:protein:protein:DNA interactions caused by the bending of DNA would also be a possible explanation for introducing a similar strategy; however, there is still not enough information for computational modelling of DNA-bending (i.e. there are not yet any computational strategies which can predict when two transcription factors which are bound to DNA with a long distance between them would have direct physical contact as a consequence of DNA bending). In addition to that, another important implication for the prediction of CRM or cis-motifs is the overlap between transcription factors which have binding sites close to each other. Based on our collision detection results, we realized that sometimes when transcription factors bind to the different grooves of DNA (major and minor) their binding sites can overlap a lot, but from a 3D point of view there is no physical overlap between factors. On the other hand, if two transcription factors bind to the same groove (usually major) then there can be a large overlap between them from a 3D point of view if there is a large overlap between their binding sites (i.e. this situation is not possible). In other words, if care is taken about the structural classification of transcription factors (i.e. if they bind to the major or minor groove) this information can also be used for CRM or cis-motif predictions.

It is interesting to note that protein-protein affinities are higher when proteins are not bound to DNA (Table 5). Interfaces between proteins that are part of a multi-complex (with DNA) can be weaker than those found in binary ones. Binding to DNA may decrease protein-protein affinities, while increasing the overall stability of the complex (significantly higher stability, student's test, p$<$0.001, Table 5). When two proteins bind freely in solution they are largely unhindered in their rotational movement so they can align themselves using the most energetically favourable orientation which gives them the optimal protein-protein binding energy. When DNA is added to the complex, the three components must arrange themselves to form a global energy minima. However the requirement of binding to DNA introduces a restriction on the possible arrangement of the components such that the protein-protein binding may be weakened by this extra strain but the additional synergistic stability of the three way complex more than compensates for this effect (Table 5).

Conclusion

It is very difficult to determine the rules governing the assembly of complexes by data-mining alone [38]. Universal conclusions for the types of complexes used are unreliable because of the limited number of available structures (44). However, many general descriptive features can be elucidated even with a modest data collection. As further structures become available, the confidence in the results presented here can be further constrained. The precedent for such studies, using similar or even smaller number of structures is well documented (e.g. [10,15,19,23]).

In this paper, we conclude that protein-protein and protein-DNA interface parameters, such as interface area, number of interface residues/atoms and hydrogen bonds, and distribution of interface residues, hydrogen bonds, van der Walls contacts and secondary structure motifs in complexes where multiple proteins are bound to DNA are no different in protein-protein, single protein-DNA or multiple proteins-DNA complexes. Thus, if we have two (or more) proteins which bind together, there will be no influence on these interface parameters. Also, if we have one protein bound to DNA, then that binding will have no influence (in terms of the interface parameters mentioned) on the types of interface interactions that can occur with subsequent protein-protein complex expansion. The water mediated contacts in interfaces of components in protein:-protein:DNA complexes play less important role (found in less quantity) in the stability and specificity of recognition then in interfaces of components in the binary protein:protein and protein:DNA complexes. Distortion is significantly higher when multiple proteins bind to DNA. This distortion is required to accommodate multiple protein binding events. The combinatorial assembly of transcription factors has been known for a long time to play an important role in stabilizing regulatory complexes. A deeper understanding of structural considerations may be helpful when predicting the assembly of transcription factor complexes. The formation of multiple protein interactions with DNA results in a decrease in protein-protein affinity and an increase in protein-DNA affinity with a net gain in overall stability for a protein-protein-DNA complex. Such effects are clearly important for modelling transcription factor cooperativity.

Materials and Methods

Definition of data sets

We selected 75 crystal complexes from the PDB database which contained two or more proteins bound to DNA with a resolution of 3.25 Å or less. We discarded all homologous complexes with less than 30% protein sequence for all protein components using the PISCES server [39,40]. Our final dataset contained 46 complexes (Table S33). We determined the UniProt ID of each protein component using the tool [41]. This dataset was called group-MultiProteins:DNA. Most of the complexes from group-Multi-Proteins:DNA are ternary (two proteins bound to DNA), but a few

of them are quaternary (three proteins bound to DNA). A very few of them contain one protein which does not make contact with DNA but is bound to another protein which does have a direct contact with DNA. We created a second dataset (group-SubMultiProteins:DNA) from group-MultiProteins:DNA which consisted of 91 structures (this number is smaller than 92, because some of the proteins do not have direct contact with DNA), each of which was a sub-structure containing only one protein unit plus DNA. In addition, we analysed a set (group-SingleProtien:DNA, Table S34) of single protein-DNA complexes (102 structures), which was a subset derived from a previous study [16]. We found 17 PDB structures (group-SingleSameProtein:DNA, Table S35) which contained single proteins and DNA, but the proteins were all components of complexes in group-MultiProteins:DNA. Corresponding subgroup of group-MultiProteins:DNA which contains complexes for each where there is a partner in the SingleSameProtein:DNA group we call this group-SubSetMultiProteins:DNA (Table S36). The group-Protein:Protein (Table S37), which contained 70 protein-protein complexes, came from a previous study [9].

Physical and chemical analysis of interfaces

We used the PISA service from the European Bioinformatics Institute [25,26] to calculate interface areas and compositions. There are two possibilities for defining the interface between two macromolecular components: the first approach defines the interface as the protein surface area which becomes inaccessible to solvents when two chains come into contact; the second method defines the interface as the set of atoms, where the atom centers from different proteins lie within a distance of 1–5 Å. Both approaches are widely used in macromolecular complex analysis and produce roughly equivalent results. The PISA service uses the first approach. The interface area between macromolecular components M1 and M2 is calculated as the difference in total accessible surface areas of isolated and interfacing structures divided by two, i.e.:

$$IA(M_1,M_2)=\frac{ASA(M_1)+ASA(M_2)-ASA(M_1,M_2)}{2} \quad (2)$$

where ASA(M1) and ASA(M2) are the accessible surface areas of macromolecular components M1 and M2 respectively, and ASA(M1M2) is the accessible surface area of the complex of M1 and M2.

We also used the PISA service to calculate hydrogen bonds, salt bridges, disulphide bonds and interface residues. However, PISA provides no information about van der Waals contacts between atoms (residues) because they may be in contact with several other residues. This is the principal difference between the outputs for van der Waals and hydrogen bonds, where inter-atomic links are well determined. However, in order to produce results comparable with previous studies, we have calculated van der Waals contacts in the following way: all atoms not involved in hydrogen bonds but separated by 3.9 Å or less are considered to be interacting through van der Waals contacts [18]. We also analyzed the statistical distribution of amino acid-amino acid and amino acid-nucleotide pairs (''interaction matrices'') for hydrogen bonds and van der Waal contacts. For all amino acid-amino acid and amino acid-nucleotide pairs we calculated contingency tables. The expected values for these tables are based on an assumption of random interactions. We evaluated the contingency tables using Fisher's exact test for count data with simulated p-values based on 200000 repetitions (GNU R). The p-value obtained by Fisher's exact test indicates whether rows and columns in contingency tables are independent or not. However, this does not provide information about which of the pairings are different from expected. To calculate this we performed individual Fisher's tests (GNU R) for each pair.

In order to determine the chemical characteristics of the interfaces, we classified the interface residues using Eisenberg's hydrophobicity scale [42] in a similar way to Lejeune et al. [16]: amino acids are assigned to groups which contain those that are positively charged (Arg and Lys), negatively charged (Asp and Glu), polar (Asn, Gln, His, Ser, and Thr), aliphatic (Ala, Ile, Leu, Met and Val), aromatic (Phe, Trp, and Tyr), and particular (Cys, Gly, and Pro). Multinomial distributions obtained in this study were compared using the Chi-square multinomial goodness-of-fit test.

In addition, a general indication of the hydrophobicity of the interfaces can be estimated using the residue interface propensities. The residue interface propensities give a measure of the relative importance of different amino acid (nucleic acid) residues in all the interfaces of complexes. The propensity values can be calculated using the accessible surface area of residues, as was done by Ellis et al. [10], or using the frequencies of residues, as was done by Lejeune et al. [16]. Both approaches have the same goal, to determine the relative importance of the different residues. Because of its simplicity, we have used the approach described in [16]. Following that, the propensity P_x for the interface residues x (x and y are amino acid or DNA structures) can be calculated by:

$$P_x=\frac{I_x \Big/ \sum_y I_y}{T_x \Big/ \sum_y T_y} \quad (3)$$

where I_x is the total number of residues x in the interface area, T_x is the total number of residues in the whole dataset and similar for T_y and I_y. If $P_x>1$ it indicates that the residue x is ''favoured'' and occurs more frequently at interfaces than in the dataset as a whole. If $P_x<1$ then residue x is ''disfavoured'' at interaction sites; in all other cases we can say that residue x is neither over- nor under-represented in the interface region in the complexes. In order to evaluate whether a particular propensity value was significantly different from 1 (either above or below), a statistical bootstrapping method was implemented similar to [10].

Structural analysis of interfaces

We analyzed the types of secondary structures present within protein-protein and protein-DNA interfaces using the PROMOTIF program [27]. PROMOTIF defines 11 different secondary structure motifs: β-turns, γ-turns, β-bulges, α-helices, 3_{10}-helices, β-strands, β-sheets, βαβ units, ψ-loop, β-hairpins, and disulphide bridges. For each structural motif we calculated propensities in the same way as we did for residue propensities (formula (3)).

Analysis of DNA distortion

DNA distortions were estimated by calculating the root-mean-square deviation (rmsd) when each DNA structure from a complex was fitted onto the corresponding canonical A-DNA and B-DNA structures as in [15], using the whole DNA from crystal strucutres and without normalization to the length of the DNA used. (Regions which are not in interactions do not have significant deformation therefore their contributions to RMSD is not big.) Canonical A-DNA and B-DNA for the nucleotide sequence (with the same length) from the complex were constructed using

X3DNA [28]. The fitting was performed with the McLachlan algorithm [43] as implemented in the program ProFit [44].

Analysis of water molecules in protein-protein and protein-DNA interactions

Water molecules are defined as interface water molecules if they are less than 3.5 Å from the atoms of the two components of a complex, as in [21]. This analysis was restricted to those structures with 2.4 Å or better resolution as the identification of water in the electron density map may be ambiguous at lower resolutions [21].

Analysis of energetic properties of interfaces

The chemical stability of complexes was analysed by calculating the free energy barrier of assembly dissociation (ΔG^{diss}) and the solvation free energy gain upon formation of the assembly (ΔG^{int}) in kJ/mol using PISA. Assemblies with higher positive values of ΔG^{diss} are more thermodynamically stable, and that value indicates that an external driving force is required to dissociate the assembly. For the calculation of ΔG^{int} and ΔG^{diss} we used structures from all six groups (-MultiProteins:DNA, -SubMultiProteins:DNA, -SingleProtein:DNA, -SingleSameProtein:DNA, -SubSetMultiProteins:DNA and –Protein:Protein).

We calculated Z-scores for intermolecular and intramolecular readouts using a ReadOut server [29]. Direct readouts (direct contacts between amino acids and base pairs) and water-mediated contacts are intramolecular energies, whereas indirect energies quantify sequence-dependent DNA conformational energies. The specificity of the complex is given by the Z-score, and larger negative values correspond to higher specificities [45]. For the calculation of the Z-score, we used the data from groups –MultiProteins:DNA, -SubMultiProteins:DNA, -SingleProteins:DNA, -SingleSameProtein, -SubSetMultiProteins:DNA.

We calculated binding energy affinities (protein-DNA) for each structure in groups –MultiProteins:DNA, -SubMultiProteins:DNA, -SingleProtein:DNA, -SingleSameProtein:DNA, and –SubSetMultiProteins:DNA using the DFIRE energy function [30].

We compared the mean of ΔG^{int}, ΔG^{diss}, the Z-score for direct and indirect readouts, and the binding energy affinities between group-MultiProteins:DNA and each of the other three groups (-SubMultiProteins:DNA, -SingleProtein:DNA and –SingleSameProtein:DNA) using student's t-test (one-tailed). Differences in the variances of corresponding values between groups were calculated using Bartlett's test. In those cases where we had significant differences in variance between groups, we used student's t-test with unequal variance.

For protein-protein complexes (group-Protein:Protein) we calculated ΔG^{int} and ΔG^{diss} using the PISA server. We have calculated protein-protein binding energy affinities for complexes from group-Protein:Protein and protein-protein subcomplexes from group-MultiProteins:DNA using DCOMPLEX [31]. We also compared the average protein-protein binding affinities, average values of ΔG^{int} and ΔG^{diss} between groups –MultiProteins:DNA and –Protein:Protein.

Collision detections and overlapping volume of two macromolecules

We calculated the number of atoms in collision and the volume of the overlapping region for protein-protein and protein-DNA interfaces from groups –MutliProteins:DNA, -SubMultiProteins:DNA, -SingleProtein:DNA and –SingleSameProtein:DNA. Collision detection between two macromolecules is actually collision detection between complex objects, where these objects are composed of collections of spheres. The most straightforward algorithm for modelling this problem (in the case of two objects: A1 and A2) is checking each sphere from object A1 against each sphere from object A2, and we know that objects A1 and A2 intersect only if one or more of these pairs intersect. For two objects with M and N spheres this algorithm requires O(MN) time to complete. There are several geometric algorithms with better speed for collision detection between objects in 3D space such as those based on bounding-volume (BV) hierarchies [46,47], algorithms based on axis-aligned bounding boxes AABB [48,49], algorithms based on oriented bounding boxes [50], and spatial hashing [51,52]. In this study we used an algorithm for collision detection based on spatial hashing [51] and axis-aligned bounding boxes AABB [48,49]. To perform this, we executed the following steps (Figure S7):

i. Make an AABB around each macromolecule.
ii. Check if any pair of AABBs overlaps. In order for two AABBs to overlap they must overlap on all three special axes. If there is no overlap then they cannot be in collision. Otherwise they may be in collision.
iii. Perform a special hashing on the overlapping region of each pair of AABBs that contain macromolecules that may be in collision.

The overlapping region (a rectangular prism) is divided into a three dimensional grid of cells. Each cell in the grid is a cube with side lengths equal to the diameter of the largest sphere (atom) in the macromolecule. This is a uniform spatial subdivision. Each sphere (atom) in the macromolecule can be assigned to the cell in which it lies using a hash function as follows: First it is necessary to make an AABB for each sphere. Then the (x,y,z) coordinates of the six side centers are assigned to their corresponding cells using the hash function (Figure 3).

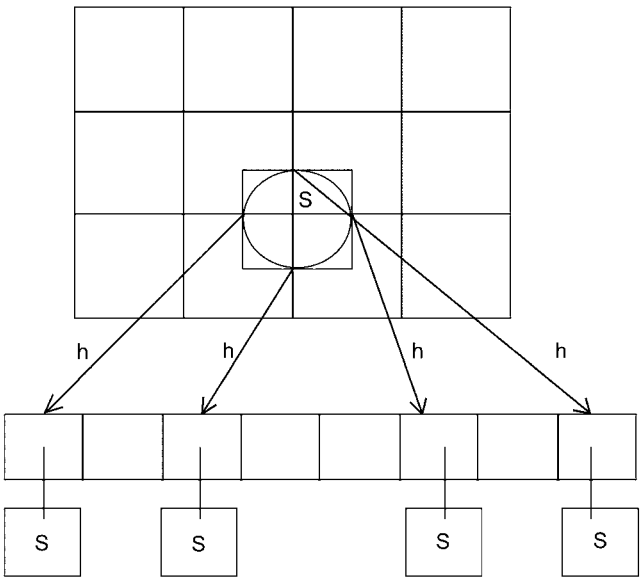


Figure 3. Assignment of hash values to the atoms of a macromolecule. Hash values are computed for all the grid cells covered by the AABB of the sphere (atom) from a macromolecule. In this case, sphere S falls into four cells and they are mapped onto a hash table.
doi:10.1371/journal.pone.0003243.g003

The hash function we used is given in formula (4) [52]:

$$h(x, y, z) = (\text{trunc}(x/l) * p1 \quad \text{xor} \quad \text{trunc}(y/l) * p2 \quad \text{xor} \quad \text{trunc}(z/l) * p3) \bmod n \tag{4}$$

where p1, p2, and p3 are large prime numbers (in our case 73856093, 19349663 and 83492791 respectively). The size of a cell is defined as 1, the hash table has a size "n". The function "trunc(x)" rounds the real number "x" down to the next integer. The function "xor" is a Boolean exclusive-or operation.

To test whether a sphere "S" from another macromolecule intersects with the first macromolecule, it suffices to find out if that sphere intersects any of the spheres of another macromolecule that share a cell with "S". The time complexity of this algorithm is linear "O(n)", where "n" is the number of sphere-atoms found in the overlapping region between two macromolecules AABBs.

We extended the collision detection algorithm so that it is able to calculate the number of atoms which are in collision and their overlapping volume. Instead of stopping the analysis as soon as two atoms are found to be in collision, the algorithm is continued until all of the atoms from the different macromolecules have been counted. From this it is a simple matter to estimate the overlapping volume from the colliding spheres.

Web-base implementation of the algorithm is freely available from http://promoterplot.fmi.ch/Collision1/. The user submits pdb files and then specifies which chains to test for collision. The output lists the number of atoms from each protein which are in collision and the volume of overlapping region. In addition, with this tool user may display 3D complex from PDB files as interactive web pages using the Corotna VRML Client plug-in or any other VRML plug-in.

Supporting Information

Figure S1 Distribution of H-bonds according to the nucleotide part (group-MultiProteins:DNA).
Found at: doi:10.1371/journal.pone.0003243.s001 (0.91 MB TIF)

Figure S2 Distribution of amino acids involved in H-bonds in protein-protein and protein-DNA interfaces (group-MultiProteins:DNA).
Found at: doi:10.1371/journal.pone.0003243.s002 (0.93 MB TIF)

Figure S3 Distribution of H-bonds according to the nucleotide part (group-SingleSameProtein:DNA).
Found at: doi:10.1371/journal.pone.0003243.s003 (0.91 MB TIF)

Figure S4 Distribution of H-bonds according to the nucleotide part (group-SubSetMultiProteins:DNA).
Found at: doi:10.1371/journal.pone.0003243.s004 (0.91 MB TIF)

Figure S5 Amino acid propensities for protein-protein and DNA-protein interfaces (group MultiProteins:DNA). Propensity values which are significantly different from 1 (either above or below), as evaluated using the statistical bootstrapping method, are marked with "*".
Found at: doi:10.1371/journal.pone.0003243.s005 (1.08 MB TIF)

Figure S6 Distribution of amino acids involved in interaction sites of protein-protein and DNA-protein (group-MultiProteins:DNA).
Found at: doi:10.1371/journal.pone.0003243.s006 (1.07 MB TIF)

Figure S7 Visualization of first several steps of the collision detection algorithm. Situation (A) represents scenario when there is on overlapping between two macromolecules and corresponding axis-aligned bounding boxes either; situation (B) represents scenario when there is no overlapping between two macromolecules but with overlapping between corresponding axis-aligned bounding boxes; situation (C) represents scenario when there is overlapping between two macromolecules and corresponding axis-aligned bounding boxes.
Found at: doi:10.1371/journal.pone.0003243.s007 (3.00 MB TIF)

Table S1 Detailed list of interface parameters for each complex from group-MultiProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s008 (0.09 MB PDF)

Table S2 The number of observed hydrogen bonds between amino acid and nucleotide moieties in protein-DNA interfaces (group-MultiProteins:DNA)
Found at: doi:10.1371/journal.pone.0003243.s009 (0.07 MB DOC)

Table S3 The number of observed hydrogen bonds between amino acid and nucleotide moieties in protein-DNA interfaces (group-SingleSameProtein:DNA)
Found at: doi:10.1371/journal.pone.0003243.s010 (0.07 MB DOC)

Table S4 The number of observed hydrogen bonds between amino acid and nucleotide moieties in protein-DNA interfaces (group-SubSetMultiProteins:DNA).
Found at: doi:10.1371/journal.pone.0003243.s011 (0.06 MB DOC)

Table S5 Number of observed van der Waals contacts between amino acid and nucleotide moieties in protein-DNA interfaces (group-MultiProteins:DNA).
Found at: doi:10.1371/journal.pone.0003243.s012 (0.06 MB DOC)

Table S6 Number of observed van der Waals contacts between amino acid and nucleotide moieties in protein-DNA interfaces (group-SingleSameProtein:DNA).
Found at: doi:10.1371/journal.pone.0003243.s013 (0.07 MB DOC)

Table S7 Number of observed van der Waals contacts between amino acid and nucleotide moieties in protein-DNA interfaces (group-SubSetMultiProteins:DNA).
Found at: doi:10.1371/journal.pone.0003243.s014 (0.06 MB DOC)

Table S8 The number of water-mediated contacts in protein-protein and protein-DNA intrerfaces of selected complexes in group-MultipleProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s015 (0.04 MB PDF)

Table S9 Detailed list of rmsd values calculated from fitting each DNA structure in the complexes from group-MultiProteins:DNA to a corresponding canonical A-DNA and B-DNA.
Found at: doi:10.1371/journal.pone.0003243.s016 (0.04 MB PDF)

Table S10 Detailed list of rmsd values calculated from fitting each DNA structure in the complexes from group-SingleProtein:DNA to a corresponding canonical A-DNA and B-DNA.
Found at: doi:10.1371/journal.pone.0003243.s017 (0.04 MB PDF)

Table S11 Detailed list of rmsd values calculated from fitting each DNA structure in the complexes from group-SingleSameProtein:DNA to a corresponding canonical A-DNA and B-DNA.

Found at: doi:10.1371/journal.pone.0003243.s018 (0.03 MB PDF)

Table S12 Detailed list of rmsd values calculated from fitting each DNA structure in the complexes from group-SubSetMutli-Proteins:DNA to a corresponding canonical A-DNA and B-DNA.
Found at: doi:10.1371/journal.pone.0003243.s019 (0.04 MB PDF)

Table S13 Average rmsd values calculated from fitting each DNA structure in the complexes from group -SubSetMultiProteins:DNA and -SingleSameProtein:DNA to a corresponding canonical A-DNA and B-DNA.
Found at: doi:10.1371/journal.pone.0003243.s020 (0.03 MB DOC)

Table S14 Detailed list of energies for each complex in group-MultiProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s021 (0.04 MB PDF)

Table S15 Detailed list of energies for each complex in group-SubMultiProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s022 (0.04 MB PDF)

Table S16 Detailed list of energies for each complex in group-SingleProtein:DNA
Found at: doi:10.1371/journal.pone.0003243.s023 (0.04 MB PDF)

Table S17 Detailed list of energies for each complex in group-SingleSameProtein:DNA
Found at: doi:10.1371/journal.pone.0003243.s024 (0.04 MB PDF)

Table S18 Detailed list of energies for each complex in group-SubSetMultiProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s025 (0.04 MB PDF)

Table S19 Detailed list of energies Z-scores (direct and indirect readouts) for each complex in group-MultiProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s026 (0.04 MB PDF)

Table S20 Detailed list of energies Z-scores (direct and indirect readouts) for each complex in group-SubMultiProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s027 (0.04 MB PDF)

Table S21 Detailed list of energies Z-scores (direct and indirect readouts) for each complex in group-SingleProtein:DNA
Found at: doi:10.1371/journal.pone.0003243.s028 (0.04 MB PDF)

Table S22 Detailed list of energy Z-scores (direct and indirect readouts) for each complex in group-SingleSameProtein:DNA
Found at: doi:10.1371/journal.pone.0003243.s029 (0.04 MB PDF)

Table S23 Detailed list of energy Z-scores (direct and indirect readouts) for each complex in group-SubSetMultiProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s030 (0.04 MB PDF)

Table S24 Detailed list of protein-DNA energy binding affinity, overlapping volume and number of atoms in collision for each complex in group-MultiProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s031 (0.04 MB PDF)

Table S25 Detailed list of protein-DNA energy binding affinity, overlapping volume and number of atoms in collision for each complex in group-SubMultiProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s032 (0.05 MB PDF)

Table S26 Detailed list of protein-DNA energy binding affinity, overlapping volume and number of atoms in collision for each complex in group-SingleProtein:DNA
Found at: doi:10.1371/journal.pone.0003243.s033 (0.05 MB PDF)

Table S27 Detailed list of protein-DNA energy binding affinity, overlapping volume and number of atoms in collision for each complex in group-SingleSameProtein:DNA
Found at: doi:10.1371/journal.pone.0003243.s034 (0.04 MB PDF)

Table S28 Detailed list of protein-DNA energy binding affinity, overlapping volume and number of atoms in collision for each complex in group-SubSetMultiProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s035 (0.04 MB PDF)

Table S29 Detailed list of protein-protein binding free energy for each protein-proteincomplex in group-MultiProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s036 (0.04 MB PDF)

Table S30 Detailed list of protein-protein binding free energy for each protein-proteincomplex in group-Protein:Protein
Found at: doi:10.1371/journal.pone.0003243.s037 (0.06 MB PDF)

Table S31 Average solvation energy (kJ/mol), free energy barrier of assembly dissociation (kJ/mol), and energy Z-scores for direct and indirect readouts for groups -SubSetMultiProteins:DNA, -SingleSameProtein:DNA
Found at: doi:10.1371/journal.pone.0003243.s038 (0.03 MB DOC)

Table S32 Average protein-DNA energy binding affinity (kJ/mol), interface overlapping volume (Å3) and average number of interface collision atoms for groups -SubSetMultiProteins:DNA, -SingleSameProtein:DNA
Found at: doi:10.1371/journal.pone.0003243.s039 (0.03 MB DOC)

Table S33 List of PDB IDs used in the study (group-Multi-Proteins:DNA), with description of component (including Swiss Prot ID) and biological process of components.
Found at: doi:10.1371/journal.pone.0003243.s040 (0.08 MB DOC)

Table S34 The list of PDB codes of complexes from group-SingleProtein:DNA
Found at: doi:10.1371/journal.pone.0003243.s041 (0.03 MB DOC)

Table S35 The list of PDB codes of complexes from group-SingleSameProtein:DNA
Found at: doi:10.1371/journal.pone.0003243.s042 (0.03 MB DOC)

Table S36 The list of PDB codes of complexes from group-SubSetMultiProteins:DNA
Found at: doi:10.1371/journal.pone.0003243.s043 (0.03 MB DOC)

Table S37 The list of PDB codes of complexes from group-Protein:Protein

Found at: doi:10.1371/journal.pone.0003243.s044 (0.03 MB PDF)

Acknowledgments

We would like to thank Prof. Torsten Schwede, Prof. Andreas Engel, Prof. Olga Mayans and Dr. Eugene Krissinel for useful discussions and Sara Oakeley for proofreading this manuscript.

Author Contributions

Conceived and designed the experiments: AT. Performed the experiments: AT. Analyzed the data: AT. Contributed reagents/materials/analysis tools: AT. Wrote the paper: AT. Participated in the design of the study, discussion of the results and drafting of the manuscript: EJO.

References

1. Sinha S, Adler AS, Field Y, Chang HY, Segal E (2008) Systematic functional characterization of cis-regulatory motifs in human core promoters. Genome Res 18: 477–488.
2. Zhao G, Schriefer LA, Stormo GD (2007) Identification of muscle-specific regulatory modules in Caenorhabditis elegans. Genome Res 17: 348–357.
3. Yu X, Lin J, Zack DJ, Qian J (2006) Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. Nucleic Acids Res 34: 4925–4936.
4. Moorman C, Sun LV, Wang J, de Wit E, Talhout W, et al. (2006) Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. Proc Natl Acad Sci U S A 103: 12027–12032.
5. Banerjee N, Zhang MQ (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. Nucleic Acids Res 31: 7024–7031.
6. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, et al. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A 99: 757–762.
7. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, et al. (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol 5: R61.
8. Cho KI, Lee K, Lee KH, Kim D, Lee D (2006) Specificity of molecular interactions in transient protein-protein interaction interfaces. Proteins 65: 593–606.
9. Chakrabarti P, Janin J (2002) Dissecting protein-protein recognition sites. Proteins 47: 334–343.
10. Ellis JJ, Broom M, Jones S (2006) Protein-RNA interactions: Structural analysis and functional classes. Proteins 66: 903–911.
11. Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM (2001) Protein-RNA interactions: a structural analysis. Nucleic Acids Res 29: 943–954.
12. Jones S, Thornton JM (1995) Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 63: 31–65.
13. Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proc Natl Acad Sci U S A 93: 13–20.
14. Jones S, Thornton JM (1997) Analysis of protein-protein interaction sites using surface patches. J Mol Biol 272: 121–132.
15. Jones S, van Heyningen P, Berman HM, Thornton JM (1999) Protein-DNA interactions: A structural analysis. J Mol Biol 287: 877–896.
16. Lejeune D, Delsaux N, Charloteaux B, Thomas A, Brasseur R (2005) Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. Proteins 61: 258–271.
17. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. J Mol Biol 285: 2177–2198.
18. Luscombe NM, Laskowski RA, Thornton JM (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res 29: 2860–2874.
19. Mandel-Gutfreund Y, Schueler O, Margalit H (1995) Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. J Mol Biol 253: 370–382.
20. Mirny LA, Gelfand MS (2002) Structural analysis of conserved base pairs in protein-DNA complexes. Nucleic Acids Res 30: 1704–1711.
21. Nadassy K, Wodak SJ, Janin J (1999) Structural features of protein-nucleic acid recognition sites. Biochemistry 38: 1999–2017.
22. Pabo CO, Nekludova L (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? J Mol Biol 301: 597–624.
23. Treger M, Westhof E (2001) Statistical analysis of atomic contacts at RNA-protein interfaces. J Mol Recognit 14: 199–214.
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
25. Krissinel E, Henrick K (2005) Detection of Protein Assemblies in Crystals. In: Berhold MRea, ed. Computational Life Sciences. Heidelberg: Springer Berlin. pp 163–174.
26. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. Journal of Molecular Biology; doi: 10.1016/j.jmb.2007.05.022.
27. Hutchinson EG, Thornton JM (1996) PROMOTIF–a program to identify and analyze structural motifs in proteins. Protein Sci 5: 212–220.
28. Lu XJ, Olson WK (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res 31: 5108–5121.
29. Ahmad S, Kono H, Arauzo-Bravo MJ, Sarai A (2006) ReadOut: structure-based calculation of direct and indirect readout energies and specificities for protein-DNA recognition. Nucleic Acids Res 34: W124–127.
30. Zhang C, Liu S, Zhu Q, Zhou Y (2005) A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. J Med Chem 48: 2325–2335.
31. Liu S, Zhang C, Zhou H, Zhou Y (2004) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. Proteins 56: 93–101.
32. Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-DNA complexes. Genome Biol 1: REVIEWS001.
33. Guharoy M, Chakrabarti P (2007) Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions. Bioinformatics.
34. Jayaram B, Jain T (2004) The role of water in protein-DNA recognition. Annu Rev Biophys Biomol Struct 33: 343–361.
35. Reddy CK, Das A, Jayaram B (2001) Do water molecules mediate protein-DNA recognition? J Mol Biol 314: 619–632.
36. Janin J (1999) Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. Structure 7: R277–279.
37. Williamson JR (2008) Cooperativity in macromolecular assembly. Nat Chem Biol 4: 458–465.
38. Sarai A, Kono H (2005) Protein-DNA recognition patterns and predictions. Annu Rev Biophys Biomol Struct 34: 379–398.
39. Wang G, Dunbrack RL Jr (2003) PISCES: a protein sequence culling server. Bioinformatics 19: 1589–1591.
40. Wang G, Dunbrack RL Jr (2005) PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 33: W94–98.
41. Martin AC (2005) Mapping PDB chains to UniProtKB entries. Bioinformatics 21: 4297–4301.
42. Eisenberg D, Weiss RM, Terwilliger TC (1982) The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature 299: 371–374.
43. McLachlan AD (1982) Rapid comparison of protein structres. Acta Crystallographica 38: 871–873.
44. Martin ACR http://www.bioinf.org.uk/software/profit/.
45. Michael Gromiha M, Siebers JG, Selvaraj S, Kono H, Sarai A (2004) Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. J Mol Biol 337: 285–294.
46. Barequet G, Chazelle B, Guibas L, Mitchell J, Tal A (1996) BOXTREE: A Hierarchical representation for Surface in 3D.
47. Hubbard P (1996) Approximation Polyhedra with Spheres for Time-critical Collision Detection. ACM trans Computer Graphics 15: 179–210.
48. Bergen G (1997) Efficient collision detection of complex deformable models using AABB trees. Journal of Graphics Tools 2: 1–13.
49. Hughes M, DiMattia C, Lin M, Manocha D (1996) Efficient and accurate interference detection for polynomial deformation and soft object animation.
50. Gottschalk S, Lin M, Manocha D (1996) OBB-tree: A hierarchical structure for rapid interference detection.
51. Turk G (1989) Interactive Collision Detection for Molecular Graphics. Chapel Hill: The University of North Carolina.
52. Teschner M, Heidelberger B, Mueller M, Romeranets D, Gross D (2003) Optimized Spatial Hashing for Collision Detection of Deformable Objects. Munich, Germany.

3.1 Supplementary information

for the paper "Computational Structural Analysis: Multiple Proteins Bound to DNA

Available on-line:
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2532747/?tool=pubmed#s4

Web-base implementation of the algorithm "Collision detections and overlapping volume of two macromolecules" is freely available from
http://promoterplot.fmi.ch/Collision1/.

4. Dependencies between transcription factors in the human, mouse and rat genome (paper IV)

Experimental strategies for detecting groups of transcription factors which work together at the level of the whole genome are very limited (too expensive, require lot of time, all resources are not available, such as antibodies from ChIP-chip). Thus, there is a need for the computational detection of transcription factor site dependencies.

In this chapter, a computational analysis of transcription factor site dependencies in human, mouse and rat genomes was performed. The results from the previous two chapters are integrated in this analysis. The scoring function introduced in chapter 2 is used to predict binding sites. The structural information observed in chapter 3 is used to model cooperativities between transcription factors. In addition, this chapter demonstrates, how *in silico* work can be combined with laboratory work. An *in vivo* validation of the computational prediction of transcription start sites for three genes (ctmp-1, ngfrap, gap-43; expressed in brain) was performed.

BMC Bioinformatics

Research article

Transcription factor site dependencies in human, mouse and rat genomes

Andrija Tomovic*[1,2], Michael Stadler[1] and Edward J Oakeley[1]

Address: [1]Friedrich Miescher Institute for Biomedical Research, Novartis Research Foundation, Basel, Switzerland and [2]Modeling and Simulation, Novartis Pharma AG, Basel, Switzerland

Email: Andrija Tomovic* - andrija.tomovic@novartis.com; Michael Stadler - michael.stadler@fmi.ch; Edward J Oakeley - edward.oakeley@novartis.com

* Corresponding author

Published: 16 October 2009

BMC Bioinformatics 2009, **10**:339 doi:10.1186/1471-2105-10-339

Received: 10 March 2009
Accepted: 16 October 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/339

Abstract

Background: It is known that transcription factors frequently act together to regulate gene expression in eukaryotes. In this paper we describe a computational analysis of transcription factor site dependencies in human, mouse and rat genomes.

Results: Our approach for quantifying tendencies of transcription factor binding sites to co-occur is based on a binding site scoring function which incorporates dependencies between positions, the use of information about the structural class of each transcription factor (major/minor groove binder), and also considered the possible implications of varying GC content of the sequences. Significant tendencies (dependencies) have been detected by non-parametric statistical methodology (permutation tests). Evaluation of obtained results has been performed in several ways: reports from literature (many of the significant dependencies between transcription factors have previously been confirmed experimentally); dependencies between transcription factors are not biased due to similarities in their DNA-binding sites; the number of dependent transcription factors that belong to the same functional and structural class is significantly higher than would be expected by chance; supporting evidence from GO clustering of targeting genes. Based on dependencies between two transcription factor binding sites (second-order dependencies), it is possible to construct higher-order dependencies (networks). Moreover results about transcription factor binding sites dependencies can be used for prediction of groups of dependent transcription factors on a given promoter sequence. Our results, as well as a scanning tool for predicting groups of dependent transcription factors binding sites are available on the Internet.

Conclusion: We show that the computational analysis of transcription factor site dependencies is a valuable complement to experimental approaches for discovering transcription regulatory interactions and networks. Scanning promoter sequences with dependent groups of transcription factor binding sites improve the quality of transcription factor predictions.

Background

Transcription factors (TFs) are a major class of DNA-binding proteins and are a crucial element in the regulation of gene expression. It is well established that many transcription factors act together to regulate gene expression in eukaryotes [1]. For example, the cooperation between E2F

and NF-Y, two main regulators of cell cycle, has been described in [2,3]. A commonly used experimental method to identify interacting proteins is tandem affinity purification (TAP), as reviewed in [4]. This approach requires the expression of recombinant fusion proteins, which is laborious, may interfere with protein function and may lead to non-physiological expression levels of the studied protein. A computational detection of potential interacting transcription factors could therefore complement experimental approaches. There are many prediction tools and databases of composite motifs and cis-regulatory modules (multiple transcription factor binding sites in a strict order and spacing) [5-21]. Most of the tools for predicting cis-regulatory modules have been limited by rigid assumptions on the architecture of the module, such as length, number and order of contained cis-motifs, distance between cis-motifs, and the DNA strand on which a binding site must appear. It has been shown that 98% of 375 known vertebrate composite elements have a distance of less than 100 bp [22]. Although these assumptions can be valid for the detection of cis-regulatory modules, they are too restrictive to allow sensitive detection of binding sites dependencies. Transcription factor cooperativity can be achieved with different spatial arrangements on different promoters. There are, for example, transcription factors which co-occur and bind to the promoter at very large distances (>1 Kbp) between them (such as GAGA and Gal4 [23]). In order to overcome these restrictions, we investigated transcription factor binding sites dependencies in terms of how often their predicted binding sites are found together within a window extending 1.5 Kb 5' and 200 bp 3' of the putative starts of transcription in human, mouse and rat genes, without any further assumption on their binding characteristics. This leads to an approach that differs from prior approaches for detecting cis-regulatory modules. Dependencies between transcription factor binding sites are evaluated using only co-occurrences among different promoter sequences, disregarding any information on arrangement and counts of occurrences within the same promoter. Binding sites of two transcription factors that appear significantly more often together (among different promoters) than expected are indicative of a dependency between them. Using this approach, even dependencies between sites that do not occur in a strictly defined order and spatial organization can be identified. Our approach for quantifying tendencies of transcription factor binding sites to co-occur is based on a scoring function which incorporates dependencies between nucleotides [24], the use of information about the structural class of each transcription factors (minor or major groove binder) and considering the possible implications of varying GC content of the sequences. The significant tendencies (dependencies) have been detected by non-parametric statistical methodology (permutation tests). Evaluation of obtained results has been performed in several ways: reports from literature (many of the significant dependencies between transcription factors have previously been confirmed experimentally); dependencies between transcription factors are not biased due to similarities in their DNA-binding sites; the number of dependent transcription factors that belong to the same functional and structural class is significantly higher than would be expected by chance; supporting evidence from GO clustering of targeting genes. The only restriction our method applies is to limit the search to the 1.7 Kb window described above, without any further restrictions on the distance between or the organization of the binding sites (cis-motifs).

Based on dependencies between two transcription factor binding sites (second-order dependencies), it is possible to construct higher-order dependencies (networks). Obtained results about dependencies among transcription factor binding sites have been further used for development of a web-based tool that allows scanning of promoter sequences for groups of dependent transcription factor binding sites http://promoterplot.fmi.ch/TFDepSSeq1/. This tool can help in predicting transcription factor binding sites in promoter analysis with relatively high sensitivity and modest specificity (which is still higher in comparison to single site prediction tools (such as [24]).

Results and Discussion

Distributions of dependencies between transcription factors

From the JASPAR database, we selected all vertebrate transcription factors (August 2007, total: 76) and made all the possible 2-order combinations (in total: $\binom{76}{2} = 2850$).

There is no comprehensive transcription factor database that would list all transcription factors with their target binding sites. From publicly available databases, JASPAR is currently the best annotated transcription factor database (new version of JASPAR database has appeared in 2008 with 88 vertebrate transcription factors). Using promoter sequences of all human, mouse and rat annotated genes (see Materials and Methods section), we analysed transcription factor site dependencies (see Material and Methods section). The total number of significant dependencies (significance level of 0.05/k, k = 75, see Methods section) in the human, mouse and rat genomes were 1438 (50.5%), 1239 (43.5%) and 1063 (37.3%), respectively [see Additional file 1]. The corresponding numbers of significant dependencies observed on background sequences [see Additional file 1] are significantly smaller (Fisher's exact test, p-value < 0.001), and are

about as high as expected based on the p-value threshold (0.05*2850). On average, the numbers of significant dependencies observed in the human, mouse and rat genomes are about four times higher than those found in the background sequences, which may indicate that statistical dependencies could correspond to real biological dependencies between transcription factors. The number of the common dependent pairs between species was also analysed [see Additional file 1] and we found a high conservation between species in terms of transcription factor dependencies, further supporting the validity of our results. Additional supporting evidence for our findings was found from the literature for many of the significant transcription factor combinations [25-29]. For example, it has been reported that SP-1 and E2F interact directly in delivering an activation signal to the basic transcription machinery [25]. In our computational analysis, dependencies between binding sites of SP-1 and E2F were detected separately in human, mouse and rat genomes, with p-values < 0.0001 in each case. There was a similar situation for USF1 and RUNX1: dependency was predicted in all three genomes, with p-values < 0.0001, and it has been reported that they interact with each other [26]. Another example is the MAX and MYC-MAX dependency which, as well as the MAX and MYCN dependency, was predicted in all three genomes, with a p-value < 0.0001, and has previously been identified [27]. The MAX-USF, MYC-USF dependency (p < 0.0001) was described in [28], NFkappaB-RELA, NFKB1-REL (p < 0.0001) in [29], and the E2F1-NFY dependency (p < 0.0001) in [2,3]. There are many other confirmatory examples which agree with the computationally predicted transcription factor dependencies. However, in order to perform a detailed investigation of the number of true and false positives we would need a precise text-mining tool to search the available scientific literature. Moreover, an additional limitation for such an investigation is that experimental information available in the literature about interacting transcription factors is certainly incomplete. Because of this, some of the results that have been evaluated as incorrect predictions (false positives) may in fact be true positives.

For each transcription factor, we analyzed the number of its dependent mates in human, mouse and rat genomes. The distributions of dependent mate numbers [see Additional file 2] are very heavily skewed from Gaussian (significantly different from Normal distributions with p-value < 0.01 detected by Kolmogorov-Smirnov, Cramer-von Mises or Anderson-Darling test for all 3 genomes) and follow a U-shaped distribution (e.g. Beta(a, b), a<1,b<1). That was expected according to the fact that there are "popular" (very often seen in dependent pairs) and "unpopular" (rarely seen in dependent pairs) transcription factors. For example a popular transcription factor in all three genomes is CREB. CREB was found to regulate ~4000 target genes in the human genome, and a majority of these are occupied in vivo [30]. In addition, there is a large number of CREB-occupied loci in the rat genome [31].

Some transcription factors, such as GATA2 and EN1, have a very high number of predicted binding sites and are thus predicted to regulate a large fraction of the analyzed promoters. For such factors, a higher number of co-occurrences with other binding sites can be observed. While our statistical approach will take this into account through an increased number of expected random co-occurrences, we wondered whether this could still cause a bias in our results. We have therefore performed a correlation analysis between the number of predicted single binding sites and the number of dependent mates for each transcription factors. We used the "Significance test for Pearson correlation" which is valid for sample sizes where N > 6 to assess these correlations. The Pearson's correlation coefficients were 0.04 (p-value = 0.75), -0.27 (p-value = 0.02) and -0.39 (p-value < 0.01) for human, rat and mouse, respectively. These results indicate that there might be reduced statistical power for factors with many predicted sites (correlation coefficient significantly different from zero in the case of rat and mouse), potentially because their lower site information content could give rise to more noise in the site predictions. However, weak correlation coefficients imply small influence of such noise on obtained results.

Similarly, we investigated the influence of binding site length on the number of dependent mates. Short binding sequences could increase the frequency of detected binding sites. We have therefore performed a correlation analysis between the length of binding sites and the number of dependent mates for each transcription factor. The Pearson's correlation coefficients were -0.30 (p-value < 0.01), -0.17 (p-value = 0.14) and -0.06 (p-value = 0.60) for human, rat and mouse, respectively. These results indicate that at least for the analysis in human, shorter binding sites tend to give rise to more dependent pairs. We cannot rule out that this is due to a higher number of false positive predictions associated to TFs with short binding sites. Yet, the observed correlation coefficients are weak, and for mouse and rat not significantly different from zero. This indicates that the resulting bias is weak and does not dominate our results.

Another potential source of bias could be the sequence composition of the promoters and binding motifs. For

example, a GC-rich promoter sequence would be more likely to contain predicted sites for GC-rich binding motifs, and detection of dependencies between corresponding factors could be biased. The stratification according to GC-content used by our resampling approach should control for the GC-content, but other compositional biases might exist that we did not account for. To investigate this issue, we performed a clustering of transcription factors based on the similarity between their binding sites [see Additional file 3]. This kind of clustering is performed in [32], and we observed [see Additional file 3] that only few TFs had sufficiently similar binding site specificities to be grouped together: the top two clusters are Cluster-15 (containing 6 transcription factors) and Cluster-5 (containing 5 transcription factors). The other clusters contain less than 5 TFs, and 32 clusters only contain a single TF. Moreover, the most popular/unpopular transcription factors (we define a popular TF as a TF which is involved in many pairwise interactions) always belong to different clusters (do not have similar binding sites), with only one exception with two popular transcription factors (ARNT, USF1). We then analyzed if dependent pairs are more likely to belong to the same cluster (Table 1). In 25 out of 469 dependent pairs (5.3%), both transcription factors are part of the same cluster. Over all possible transcription factor pairs, both factors belong to the same cluster in 33 of 1507 pairs (2.2%). This indicates that similar binding site specificity might increase the chance to be dependent by about 2.5-fold, but would still only account for a minority of predicted dependent pairs. Taken together, these results suggest that dependencies between transcription factors cannot be explained by similarity of their DNA-binding sites.

Next, we investigated how many dependent pairs contain transcription factors that belong to the same structural class, using the classification from JASPAR [33]. It has been reported that transcription factors from the same structural class tend to bind in a similar way [33-37]. We found that belonging to the same structural class is related to dependencies between transcription factors (Figure 1). This is also in agreement with the statement that similar structures imply similar functions, and similar functions imply possible transcription factor binding site dependencies. An alternative way of classifying transcription factors is based on their functions (i.e. biological processes) obtained from [38]. We investigated the distribution of dependencies according to this classification (which only covers 51 of the 76 factors used in this work), in a similar way to the structural classification. In this situation (which is more relevant for this study), we expected that transcription factors that belong to the same functional group (have the same or similar biological processes) should be dependent more often than transcription factors from the different functional class. Indeed, the number of dependent transcription factors that belong to the same functional class is significantly higher ($p = 0.04$, Chi-square test) than randomly expected in the human, rat and mouse genomes (Figure 1). For the functional analysis we did not use the all transcription factors used in this study, because for some there was no reported functional class available in [38]. This could have limited our statistical sensitivity and might be the reason why the functional enrichment was only marginally significant.

Finding groups of genes that are correlated throughout a set of experiments leads to the hypothesis that these genes are involved in common functions [39]. Further, we can expect that these genes have similar sets of dependent transcription factor binding sites. Knowledge of these sets may be crucial for further understanding of regulatory networks. Following this we investigated distributions of dependent transcription factor binding sites using the GO ontology classification (biological process and molecular function) of target genes whose promoters we used in the study, using only GO classes that contained at least 25 genes. Clustering of dependent TFs was performed in the following way: each dependent pair of TFs which had in its target list at least 80% of promoters (genes) that belong to the given GO class is assigned as relevant for that class. All results are available from http://promoterplot.fmi.ch/TFDEP1/TFdepGO.html. The predictions of dependent transcription factor binding sites are more likely to be true if they are supported by multiple lines of evidence. Figure 2 represents Venn diagrams for human, mouse and rat results separately. Venn diagrams show the number of total predicted dependent pairs, the number of predicted dependent pairs conserved in two or three species, the number of predicted dependent pairs supported by GO, and the number of predicted dependent pairs supported

Table 1: Distributions of pair dependencies according the binding sites similarity clustering.

	Dependent pairs A-B*	Independent pairs A-B*
A&B belong to the same cluster	25	8
A&B belong to the different cluster	444	1063

p-value = 7.692106e-08 (Fisher's exact test)

*transcription factors for which cluster is not assigned [see Additional file 3] are omitted from analysis

The number (percent) of dependent/independent pairs (in all there genomes human+mouse+rat intersection) that belong to the same/different cluster (clustering of transcription factors is performed based on the similarity between their binding sites, see Additional file 3).

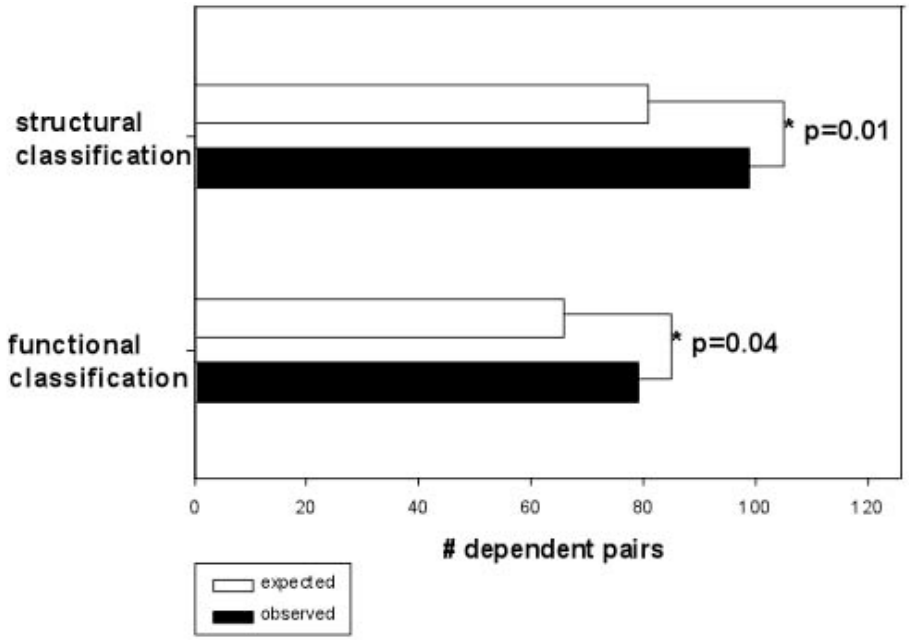


Figure 1

Distributions according to the structural and functional classification. Expected (random) and observed distributions of dependent pairs of TFs which belong to the same structural/functional class (* $p < 0.05$, Chi-square test; Expected distribution gives the numbers of dependent pairs of transcription factors which belong to the same structural/functional class that one would expect to obtain if there is no difference between proportions of dependent pairs that contain transcription factors from the same and different structural/functional classes).

by overlapped supporting evidence. We can see that the highest number of dependent pairs is supported by 2 evidences and all dependent pairs from GO analysis are supported by other 2 evidences for each species, further supporting the validity of our results. Another potential way to investigate dependencies between transcription factors according to the GO classification of their target genes would be to group the promoters belonging to the same GO cluster and perform the same analysis (see section 2.1) as performed previously with the set of all promoters. However, in practice this approach proved underpowered because of the limited number of promoters in each GO class. There were too few promoters to apply the same re-sampling techniques used for the whole genome.

It is likely that some protein-DNA complexes not only contain two, but three or more cooperating transcription factors. In order to identify such groups of more than two dependent sites, one could apply the same method as for pairs. In practise however, it is not feasible to enumerate and analyze all combinations of three or more transcription factor binding sites (for example, there are 70300 groups of three and over 1.2 million groups of 4 factors from Jaspar). Instead, we used the results on significantly associated pairs for extrapolation. Starting from dependencies of order two, we analyzed the dependencies of higher orders as fully or partially connected transcription factor networks. To make all results easily accessible, we have provided a web-based tool, freely accessible from http://promoterplot.fmi.ch/TFDEP1/ where users can search by transcription factor name and retrieve our results on dependencies (full and partial). For stringent searching, users can require the transcription factor network to be fully connected (e.g. for A-B-C dependencies it is necessary to have A-B, A-C and B-C dependencies) and represents exactly the results which would be obtained via direct enumeration. Partial connectivity is less stringent (e.g. for third-order only two combinations are necessary to be dependent) and represents a less stringent approximation of the full enumeration results. Information obtained in this way can be useful for designing biological experiments where information about transcription factors that may cooperate is useful (design of regulatory gene networks for various processes). In addition, the results obtained about dependencies are potentially useful for better understanding transcriptional networks in human, mouse and rat genomes.

Computational prediction of groups of dependent transcription factors binding sites

Results from descriptive data-mining about dependencies between transcription factor binding sites can be used for

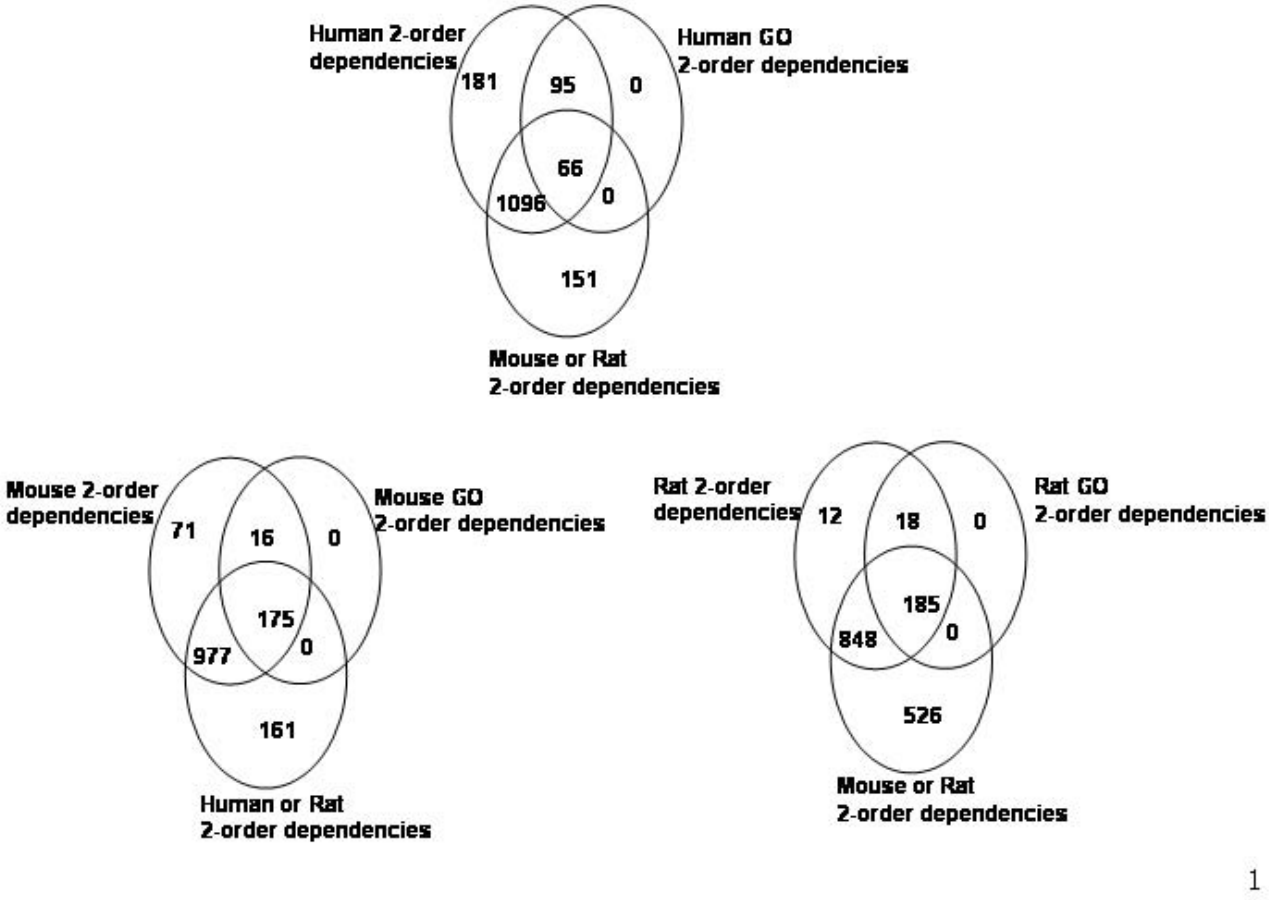


Figure 2

Venn diagrams of the number of dependent transcription factor binding sites pairs in human, mouse and rat genome. Venn diagrams show the number of total predicted dependent pairs, the number of predicted dependent pairs conserved in two or three species, the number of predicted dependent pairs supported by GO, and the number of predicted dependent pairs supported by overlapped supporting evidence.

the computational prediction of modules of dependent binding sites. In order to evaluate the proposed tool, we used experimentally verified data from [40,41]. From the dataset of transcription factors which we used in this study, we selected a subset which was known to be involved in the regulation of skeletal muscle gene expression: MEF2, SP-1, SRF, MZF1_1-4 and MZF1_5-13. It is known that a set of nine human genes (NM_184041, NM_001927, NM_002479, NM_079422, NM_003281, NM_000257, NM_002471, NM_001100 and NM_005159) is regulated by combinatorial interactions between the transcription factors listed above [41]. First,

Table 2: Computational prediction of groups of dependent transcription factors binding sites.

2-order TF dependency	# (%)of promoters where module has been detected
MZF1_5-13 ↔ SP-1	9 (100%)
MZF1_1-4 ↔ MZF1_5-13	9 (100%)
MZF_1-4 ↔ SP-1	9 (100%)
MEF2 ↔ SRF	1 (11%)

General form of output after scanning promoter sequences for the given combination of transcription factors A and B.

we noticed that based on second order dependencies (in human) among these transcription factors (Table 2) it was possible to construct fifth-order partial dependencies between them. We used the 2 Kbp upstream region of the nine human genes and scanned them with modules of order 2. We found (Table 2) that almost all second-order modules were detected in all nine promoters.

Only module MEF2-SRF was not detected in all sequences, however there are other combinations that include one of these two transcription factors detected in more sequences. This is not a surprise because not only these 5 transcription factors are involved in the regulation of skeletal muscle genes.

In order to further demonstrate the practical application of the proposed tool, we can simulate the following scenario: if we know that one specific transcription factor is involved in the regulation of a set of genes, and we would like to know which other possible transcription factors might be involved, then we could use the proposed tool to create a list of candidates. Specifically, using the set of nine genes that showed skeletal muscle expression we could start from the any of the 5 mentioned transcription factors and then find the factors that might interact with it in the regulation of these nine genes. Using the proposed tool, we were able to predict all the other known transcription factors reported to be involved in the regulation of these genes (true positives). However, we also determined another set of transcription factors for which no experimental support exists (which we might consider as potential false positives).

In order to perform more detailed validation test, we used transcription factors that were predicted and experimentally identified as true positives, transcription factors that were not predicted but experimentally reported for a given promoter as false negatives, transcription factors that were neither predicted nor experimentally reported as true negatives and transcription factors that are predicted but not experimentally reported are false positives (Table 3, with muscle specific data from [40,41]). The second order dependencies have been used in this evaluation. In addition, we have used promoters (and corresponding transcription factors: HLF, TCF1(HNF1), FOXa2 (HNF3), RORA, SOX17, cEBP, HNF4) of human liver specific genes from [42] and performed similar validation (Table 4). We noticed that sensitivity is relatively high and specificity relatively low. While our method could detect almost all true positives from both experiments, it produced many false positive predictions similar to other tools for prediction of transcription factor-binding sites. However, it is important to mention that it is not guaranteed that the experimentally reported transcription factors represent the complete set of factors for the given genes (true positives). Therefore, some of the false positives might be true positives and the actual specificity could be higher than estimated here. In comparison to single site prediction tools (such as [24], Table Sup eight-three and tools reported there), our tool has an increased specificity and sensitivity.

Conclusion

In this paper we describe a data-mining study to identify transcription factor site dependencies in the human, mouse and rat genomes. Many of the predicted dependent

Table 3: Evaluation of prediction of dependent transcription factor binding sites using transcription factors involved in the regulation of skeletal muscle gene expression.

Promoter of human gene (Gene RefSq ID)	TP	TN	FP	FN	Specificity	Sensitivity
NM_000257	3	21	50	2	0.32	0.6
NM_001100	4	24	47	1	0.35	0.8
NM_001927	4	25	46	1	0.36	0.8
NM_002471	3	25	46	2	0.37	0.6
NM_002479	4	20	51	1	0.29	0.8
NM_003281	4	22	49	1	0.32	0.8
NM_005159	5	22	49	0	0.31	1
NM_079422	4	21	50	1	0.31	0.8
NM_184041	3	27	45	1	0.38	0.75

TP-true positives, FP-false positives, TN-true negative, FN-false negative, sensitivity = TP/(TP+FN), specificity = TN/(TN+FP)

Table 4: Evaluation of prediction of dependent transcription factor binding sites using transcription factors involved in the regulation of human liver.

Promoter of human gene (Ensembl ID)	TP	TN	FP	FN	Specificity	Sensitivity
ENSG00000150526	6	23	46	1	0.33	0.857
ENSG00000017427	6	20	49	1	0.29	0.857
ENSG00000084674	6	23	46	1	0.33	0.857
ENSG00000115718	5	23	46	2	0.33	0.714
ENSG00000116833	6	28	41	1	0.41	0.857
ENSG00000126218	6	21	48	1	0.30	0.857
ENSG00000136872	6	20	49	1	0.29	0.857
ENSG00000163581	6	25	44	1	0.36	0.857
ENSG00000163631	6	21	48	1	0.30	0.857
ENSG00000167165	6	28	41	1	0.40	0.857
ENSG00000167910	6	27	42	1	0.39	0.857
ENSG00000171759	6	26	43	1	0.37	0.857
ENSG00000173531	6	23	46	1	0.33	0.857
ENSG00000180432	6	23	46	1	0.33	0.857
ENSG00000101076	6	22	47	1	0.32	0.857
ENSG00000163631	6	21	48	1	0.30	0.857
ENSG00000145321	6	23	46	1	0.33	0.857
ENSG00000169562	6	22	47	1	0.32	0.857
ENSG00000132437	6	21	48	1	0.30	0.857
ENSG00000105398	6	24	45	1	0.35	0.857
ENSG00000131482	6	25	44	1	0.36	0.857
ENSG00000198610	6	25	44	1	0.36	0.857

TP-true positives, FP-false positives, TN-true negative, FN-false negative, sensitivity = TP/(TP+FN), specificity = TN/(TN+FP)

transcription factors had been confirmed previously *in vitro* or *in vivo* and have been reported in the literature: these represent partial validation of our approach (agreement between statistical and biological/experimentally confirmed/dependencies). Dependencies between transcription factors are not biased by similarities in their DNA-binding sites. The distribution of transcription factors, whose binding sites are dependent, according to their functional classification shows that they tend to be involved in same biological process. Genes that are involved in common functions tend to have similar sets of dependent transcription factor binding sites. Knowing these sets may further our understanding of gene regulation networks. This is why we provided distributions of dependent transcription factor binding sites in GO ontology classes of target genes whose promoters we used in the study and these results are available from http://promoterplot.fmi.ch/TFDEP1/TFdepGO.html. Starting from the dependencies of order 2, it is possible to construct higher order dependencies (networks). All results can be obtained via the web tool http://promoterplot.fmi.ch/TFDEP1/. This information may help others in their inves-

tigation of transcriptional processes in human, mouse and rat. In addition, we demonstrated how the information obtained about dependencies could be used for the computational prediction of modules of dependent transcription factor binding sites http://promoterplot.fmi.ch/TFDepSSeq1/. We validated the tool using experimentally verified data set of transcription factors involved in the regulation of skeletal muscle expression. We also demonstrated how the proposed tool might be applied. Computational analysis of transcription factor site dependencies is a complement to experimental approaches for discovering transcription regulatory interactions and networks.

Methods

De novo *detection of transcription factor site dependencies*

The dataset used in this study comprised promoter sequences (1500 bp upstream to 200 bp downstream of annotated transcription start sites) of 18,799 human (Ensembl Build 40, NBCI v36, hg18), 17,954 mouse (Ensembl v38, NCBI m35, mm7) and 6,723 rat genes (Ensembl v22, NCBI v3.1, rn3) taken from the cisRED database, August 2007 [43]. The set of vertebrate transcription factors (total 76) with their binding sites was obtained from the non-redundant, curated and publically available database JASPAR [44,45] (August, 2007). We also used negative control sequences as a background in order to see how many dependent transcription factors can be found in sequences which are not real promoters of selected genes. Background sequences were generated for each species as described in [43], of 1000 concatenated search regions that were randomly selected from the genome's entire set of search regions.

In order to detect transcription factor site dependencies, we first enumerated all second-order combinations of transcription factors. Then, using the new scoring function introduced in our previous work [24], we predicted binding sites for the given combination of transcription factors on the aforementioned human/mouse/rat promoter sequences. It is difficult to define a single optimal score threshold for all TFs. Individually optimized thresholds might be necessary to account for varying degrees of specificity inherent to some TFs. Nevertheless, we used universal but distance specific thresholds for this study: 0.88 if the distance between binding sites was longer than 5 bp, otherwise 0.80, because transcription factors with direct contacts between them can make more stable complexes with DNA even though their DNA-binding affinities may be lower, as discussed in [46]. In our previous paper [24] we suggested values between 0.8 and 0.9 as optimal medium stringency thresholds for the prediction of single transcription factor binding sites. Very similar results are obtained if other thresholds are chosen from this interval, with a ~5-10% difference between them (data not shown). In addition for detection binding site dependencies, we also included information about the structural class of each transcription factor from the JASPAR database. It is known that most transcription factors bind to the major DNA groove, but some of them bind to the minor groove. Practically, this means that overlapping binding sites can be possible if one transcription factor binds to the major and other to the minor groove (acceptable structural arrangement). The strand of DNA determines the orientation of transcription factors on DNA. Based on this observation, we allow that the binding sites of two transcription factors can overlap (partially or even completely) if those two transcription factors bind to DNA in a different way (one to the major and one to the minor groove). We analyzed both strands of the promoter sequences. In summary, if there are two binding sites (of different transcription factors) are further apart than 5 bp, we treated them as "predicted" if scoring function is higher than 0.88. If the distance is shorter then 5 bp (or there is overlap between them) with acceptable structural arrangement we treated them both as "predicted" even if scoring function for any of them is smaller of 0.88 (but >= 0.8); finally if two binding sites (of different transcription factors) overlap with an unacceptable structural arrangement, then we treated only the one with the higher score as "predicted".

For each promoter sequence we calculated the CG context (%G + %C). Histogram distributions of GC content are given in Additional file 4. We employed a Monte-Carlo resampling approach to determine the significance of observed co-occurring transcription factor binding sites as follows. For a given combination of two transcription factors A and B, and the list of promoter sequences, the results of the initial predictions can be represented as a table in which we have calculated the number of promoter sequences that have binding sites for both transcription factors A and B [see Additional file 5]:

$$CountAB = \sum_{i=1}^{n} I(A_i, B_i) \quad (1)$$

where

$$I(\mathrm{A_i}, \mathrm{B_i}) = \begin{cases} 1 \text{ if } A_i = 1, B_i = 1 \\ 0, otherwise \end{cases} \quad (2)$$

and n is the total number of sequences, $\mathrm{A_i} = 1$ means that sequence i has binding sites of transcription factor A, $\mathrm{A_i} = 0$ means that sequence i has no binding sites of transcription factor A, and similar for $\mathrm{B_i}$.

Then, in a series of R replicates, we performed a permutation of the initial table [see Additional file 5] in the fol-

lowing way: for each promoter sequence i ($1 \le i \le n$), we randomly assigned to it another promoter sequence j ($1 \le j \le n$) which had a similar GC content, and we replaced (swapped) values in column A between rows (sequences) i and j (i.e. $A_i <-> B_j$).

In order to define the term "similar GC content between sequences" we could have used equal intervals of GC content. However, we noticed that this would result in a smaller number of sequences for permutation in high and low GC bins. To correct for this, we produced 50 bins with a fixed number of promoters per bin [see Additional file 6]. In this way, we ensured enough possible permutations for each sequence and its corresponding GC content. Using this method, we produced R permuted tables, and for each permuted table we counted how many times we had the value 1 in columns A and B (CountPermjAB was performed substituting "CountPermAB" for "CountAB" in equation (1)) for each table j (j = 1,... R). Finally, a p-value was calculated in the following way:

$$p - value = \frac{1 + \sum_{j=1}^{R} G(CountAB, CountPerm_j AB)}{1+R} \quad (3)$$

where

$$G(\text{CountAB}, \text{CountPerm}_j\text{AB}) = \begin{cases} 1 \text{ if CountPerm}_j AB \ge CountAB \\ 0, otherwise \end{cases} \quad (4)$$

and R is the resample size (number of replicates), and adding 1 is the pseudocount that prevents us from underestimating the p-value when it is low or zero. We used an adjusted p-value (with Bonferroni's correction) to correct for multiple testing errors. Dependencies were declared significant if the computed p-value was smaller than 0.05/k (where k is the number of multiple tests). We determine the number of re-sampling runs using the following formula:

$$1/R << P\_threshold \quad (5)$$

where P_threshold is the significance p-value threshold selected which, in our case, corresponded to P_threshold = 0.05/k where k = 75. We therefore selected R = 15,000 as a compromise between accuracy in p-value estimation and calculation time (R>>k/0.05 = 1500).

Higher-order transcription factor site dependencies

Starting from dependencies of order two, we constructed dependencies of higher orders in the following way: if transcription factors A-B, B-C and A-C are all dependent, then we can claim that there is an order three dependency between transcription factors A, B and C. (Note: it is not true if only A-B and B-C are dependent pairs but A-C is not). Third-order dependencies between the transcription factors A, B and C can be represented as fully connected graph as shown in Additional file 7. Other forms of third-order dependencies (partial third-order dependencies) of transcription factors (when any of two pairs of three transcription factors are dependent) can be represented using a not fully connected graph [see Additional file 7]. Higher order dependencies between factors can be represented in a similar way.

Scanning tool for predicting groups of dependent transcription factor binding sites

The computational prediction of cis regulatory motifs of dependent transcription factors in scanning form can be performed using information about dependencies between transcription factor binding sites using the scoring function which we introduced in a previous paper [24] and, in addition, structural information (possible position binding) between transcription factors as we described in section "De novo detection of transcription factor site dependencies". We used universal but distance specific thresholds for the scoring function as described in the same section. This method is implemented as a web-based tool and it is available from: http://promoterplot.fmi.ch/TFDepSSeq1/. Different cut-off values in the range between 0.8 and 0.9 only had a minor influence on the results in Table 3 and 4 (slightly varying only in the number of false positives for different promoters from the here shown numbers for different cut-off values). If very different cut-off values are chosen (above 0.9 or bellow 0.8), a greater impact on the results as shown in Table 3 and 4 can be observed. As indicated in the section "De novo detection of transcription factor site dependencies", we think however that it is not recommended to use such cut-off values.

Authors' contributions

AT designed the study, performed computational analysis, created supported web tool and drafted the manuscript. MS and EJO participated in the design of the study, discussion of the results and drafting of the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Distribution of dependencies of order 2 in the human, mouse and rat genomes using real promoters sequences and background sequences.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-339-S1.PDF]

Additional file 2

Distributions of number of dependent mates in human, mouse and rat genome. *File containing 3 histograms of number of dependent mates for each transcription factor in human, mouse and rat genome.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-339-S2.PDF]

Additional file 3

Distribution of dependent mates for each transcription factor in human, mouse and rat genome, including cluster information about similarity between binding sites.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-339-S3.PDF]

Additional file 4

Distribution of GC content in the human, mouse and rat promoters. *File containing 3 histograms and corresponding fitted normal distributions.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-339-S4.PDF]

Additional file 5

Scanning promoter sequences. *File containing a table that represents a general form of output after scanning promoter sequences for the given combination of transcription factors A and B.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-339-S5.PDF]

Additional file 6

Distributions of GC content in human promoters, represented by a histogram of 50 bins. *File containing 3 histograms of 50 bins each.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-339-S6.PDF]

Additional file 7

Representation of higher order dependencies between transcription factors A, B and C. *File containing fully connected graph (represents full 3-order dependencies) and not fully connected graph (represents partial 3-order dependencies).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-339-S7.PDF]

Acknowledgements

We would like to thank Anthony (Tony) Rossini (Novartis, Modeling & Simulation) for useful methodology advice. This work was supported by the Novartis Research Foundation.

References

1. GuhaThakurta D: **Computational identification of transcriptional regulatory elements in DNA sequence.** *Nucleic Acids Res* 2006, **34:**3585-3598.
2. van Ginkel PR, Hsiao KM, Schjerven H, Farnham PJ: **E2F-mediated growth regulation requires transcription factor cooperation.** *J Biol Chem* 1997, **272:**18367-18374.
3. Caretti G, Salsi V, Vecchi C, Imbriano C, Mantovani R: **Dynamic recruitment of NF-Y and histone acetyltransferases on cell-cycle promoters.** *J Biol Chem* 2003, **278:**30435-30440.
4. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Seraphin B: **The tandem affinity purification (TAP) method: a general procedure of protein complex purification.** *Methods (San Diego, Calif)* 2001, **24:**218-229.
5. Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE: **Decoding human regulatory circuits.** *Genome Res* 2004, **14:**1967-1974.
6. Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, *et al.*: **Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression.** *Genome Res* 2006, **16:**656-668.
7. Sharan R, Ben-Hur A, Loots GG, Ovcharenko I: **CREME: Cis-Regulatory Module Explorer for the human genome.** *Nucleic Acids Res* 2004, **32:**W253-W256.
8. Sharan R, Ovcharenko I, Ben-Hur A, Karp RM: **CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments.** *Bioinformatics (Oxford, England)* 2003, **19(Suppl 1):**i283-i291.
9. Choi D, Fang Y, Mathers WD: **Condition-specific coregulation with cis-regulatory motifs and modules in the mouse genome.** *Genomics* 2006, **87:**500-508.
10. Wang H, Zhang Y, Cheng Y, Zhou Y, King DC, Taylor J, Chiaromonte F, Kasturi J, Petrykowska H, Gibb B, *et al.*: **Experimental validation of predicted mammalian erythroid cis-regulatory modules.** *Genome Res* 2006, **16:**1480-1492.
11. Donaldson IJ, Gottgens B: **CoMoDis: composite motif discovery in mammalian genomes.** *Nucleic Acids Res* 2007, **35:**e1.
12. Liu CC, Lin CC, Chen WS, Chen HY, Chang PC, Chen JJ, Yang PC: **CRSD: a comprehensive web server for composite regulatory signature discovery.** *Nucleic Acids Res* 2006, **34:**W571-W577.
13. Donaldson IJ, Gottgens B: **TFBScluster web server for the identification of mammalian composite regulatory elements.** *Nucleic Acids Res* 2006, **34:**W524-W528.
14. King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC: **Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences.** *Genome Res* 2005, **15:**1051-1060.
15. Schones DE, Smith AD, Zhang MQ: **Statistical significance of cis-regulatory modules.** *BMC Bioinformatics* 2007, **8:**19.
16. Yan X, Mehan MR, Huang Y, Waterman MS, Yu PS, Zhou XJ: **A graph-based approach to systematically reconstruct human transcriptional regulatory modules.** *Bioinformatics (Oxford, England)* 2007, **23:**i577-i586.
17. Ferretti V, Poitras C, Bergeron D, Coulombe B, Robert F, Blanchette M: **PReMod: a database of genome-wide mammalian cis-regulatory module predictions.** *Nucleic Acids Res* 2007, **35:**D122-D126.
18. Jegga AG, Chen J, Gowrisankar S, Deshmukh MA, Gudivada R, Kong S, Kaimal V, Aronow BJ: **GenomeTrafac: a whole genome resource for the detection of transcription factor binding site clusters associated with conventional and microRNA encoding genes conserved between mouse and human gene orthologs.** *Nucleic Acids Res* 2007, **35:**D116-D121.
19. Zhao G, Schriefer LA, Stormo GD: **Identification of muscle-specific regulatory modules in Caenorhabditis elegans.** *Genome Res* 2007, **17:**348-357.
20. Alkema WB, Johansson O, Lagergren J, Wasserman WW: **MSCAN: identification of functional clusters of transcription factor binding sites.** *Nucleic Acids Res* 2004, **32:**W195-W198.
21. Di Cara A, Schmidt K, Hemmings BA, Oakeley EJ: **PromoterPlot: a graphical display of promoter similarities by pattern recognition.** *Nucleic Acids Res* 2005, **33:**W423-W426.
22. Klein H, Vingron M: **Using Transcription Factor Binding Site Co-Occurrence to Predict Regulatory Regions.** *Genome Informatics* 2007, **18:**109-118.
23. Mahmoudi T, Katsani KR, Verrijzer CP: **GAGA can mediate enhancer function in trans by linking two separate DNA molecules.** *The EMBO journal* 2002, **21:**1775-1781.
24. Tomovic A, Oakeley EJ: **Position dependencies in transcription factor binding sites.** *Bioinformatics (Oxford, England)* 2007, **23:**933-941.

25. Karlseder J, Rotheneder H, Wintersberger E: **Interaction of Sp1 with the growth- and cell cycle-regulated transcription factor E2F.** *Mol Cell Biol* 1996, **16:**1659-1667.
26. Carabana J, Ortigoza E, Krangel MS: **Regulation of the murine Ddelta2 promoter by upstream stimulatory factor 1, Runx1, and c-Myb.** *J Immunol* 2005, **174:**4144-4152.
27. Crouch DH, Fisher F, Clark W, Jayaraman PS, Goding CR, Gillespie DA: **Gene-regulatory properties of Myc helix-loop-helix/leucine zipper mutants: Max-dependent DNA binding and transcriptional activation in yeast correlates with transforming capacity.** *Oncogene* 1993, **8:**1849-1855.
28. Walhout AJ, Gubbels JM, Bernards R, Vliet PC van der, Timmers HT: **c-Myc/Max heterodimers bind cooperatively to the E-box sequences located in the first intron of the rat ornithine decarboxylase (ODC) gene.** *Nucleic Acids Res* 1997, **25:**1493-1501.
29. Kunsch C, Ruben SM, Rosen CA: **Selection of optimal kappa B/Rel DNA-binding motifs: interaction of both subunits of NF-kappa B with DNA is required for transcriptional activation.** *Mol Cell Biol* 1992, **12:**4412-4421.
30. Zhang X, Odom DT, Koo SH, Conkright MD, Canettieri G, Best J, Chen H, Jenner R, Herbolsheimer E, Jacobsen E, *et al.*: **Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues.** *Proc Natl Acad Sci USA* 2005, **102:**4459-4464.
31. Impey S, McCorkle SR, Cha-Molstad H, Dwyer JM, Yochum GS, Boss JM, McWeeney S, Dunn JJ, Mandel G, Goodman RH: **Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions.** *Cell* 2004, **119:**1041-1054.
32. Kielbasa SM, Gonze D, Herzel H: **Measuring similarities between transcription factor binding sites.** *BMC Bioinformatics* 2005, **6:**237.
33. Sandelin A, Wasserman WW: **Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics.** *Journal of molecular biology* 2004, **338:**207-215.
34. Morozov AV, Siggia ED: **Connecting protein structure with predictions of regulatory sites.** *Proc Natl Acad Sci USA* 2007, **104:**7068-7073.
35. Narlikar L, Gordan R, Hartemink AJ: **Nucleosome Occupancy Information Improves de novo Motif Discovery.** *RECOMB* 2007:107-121.
36. Narlikar L, Gordan R, Ohler U, Hartemink AJ: **Informative priors based on transcription factor structural class improve de novo motif discovery.** *Bioinformatics (Oxford, England)* 2006, **22:**e384-e392.
37. Mahony S, Auron PE, Benos PV: **DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies.** *PLoS Comput Biol* 2007, **3:**e61.
38. Brivanlou AH, Darnell JE Jr: **Signal transduction and the control of gene expression.** *Science* 2002, **295:**813-818.
39. Gyenesei A, Wagner U, Barkow-Oesterreicher S, Stolte E, Schlapbach R: **Mining co-regulated gene profiles for the detection of functional associations in gene expression data.** *Bioinformatics (Oxford, England)* 2007, **23:**1927-1935.
40. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *Journal of molecular biology* 1998, **278:**167-181.
41. Defrance M, Touzet H: **Predicting transcription factor binding sites using local over-representation and comparative genomics.** *BMC Bioinformatics* 2006, **7:**396.
42. Ho Sui SJ, Fulton DL, Arenillas DJ, Kwon AT, Wasserman WW: **oPOSSUM: integrated tools for analysis of regulatory motif over-representation.** *Nucleic Acids Res* 2007, **35:**W245-W252.
43. Robertson G, Bilenky M, Lin K, He A, Yuen W, Dagpinar M, Varhol R, Teague K, Griffith OL, Zhang X, *et al.*: **cisRED: a database system for genome-scale computational discovery of regulatory elements.** *Nucleic Acids Res* 2006, **34:**D68-D73.
44. Lenhard B, Wasserman WW: **TFBS: Computational framework for transcription factor binding site analysis.** *Bioinformatics (Oxford, England)* 2002, **18:**1135-1136.
45. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles.** *Nucleic Acids Res* 2004, **32:**D91-D94.
46. Tomovic A, Oakeley EJ: **Computational structural analysis: multiple proteins bound to DNA.** *Plos One* 2008, **3:**e3243.

4.1 Additional material

for the paper

Transcription factor site dependencies in the human, mouse and rat genome

A. Tomovic, M. Stadler, E.J. Oakeley

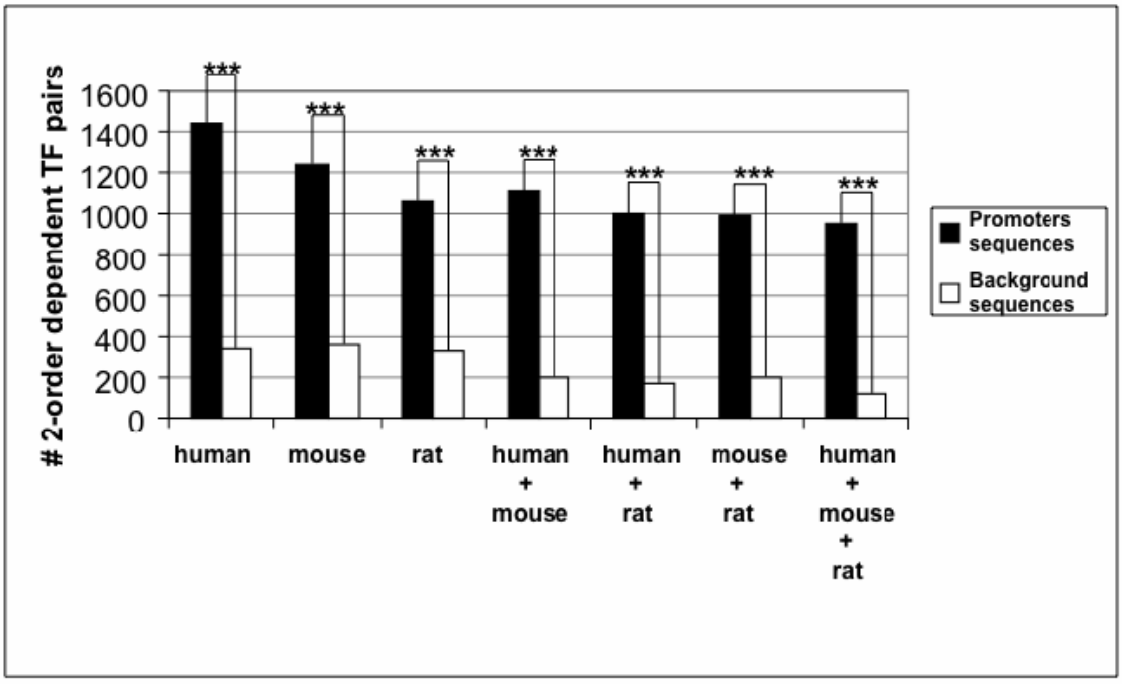


Figure A1 - Dependency distributions

Distribution of dependencies of order 2 in the human, mouse and rat genomes using real promoters sequences (black) and background sequences (white).

*** p-value<0.001 calculated by Fisher's exact test.

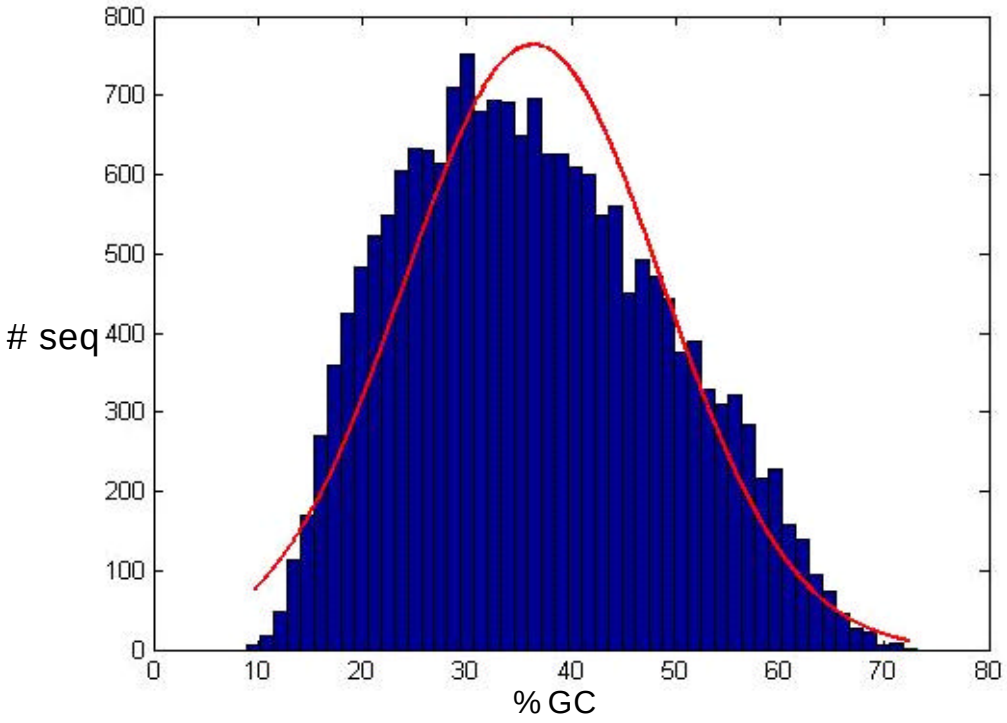


Figure A2. Distribution of GC content in human promoters. Red line represents fitted normal distributions (with mean 36.37 and standard deviation 12.39).

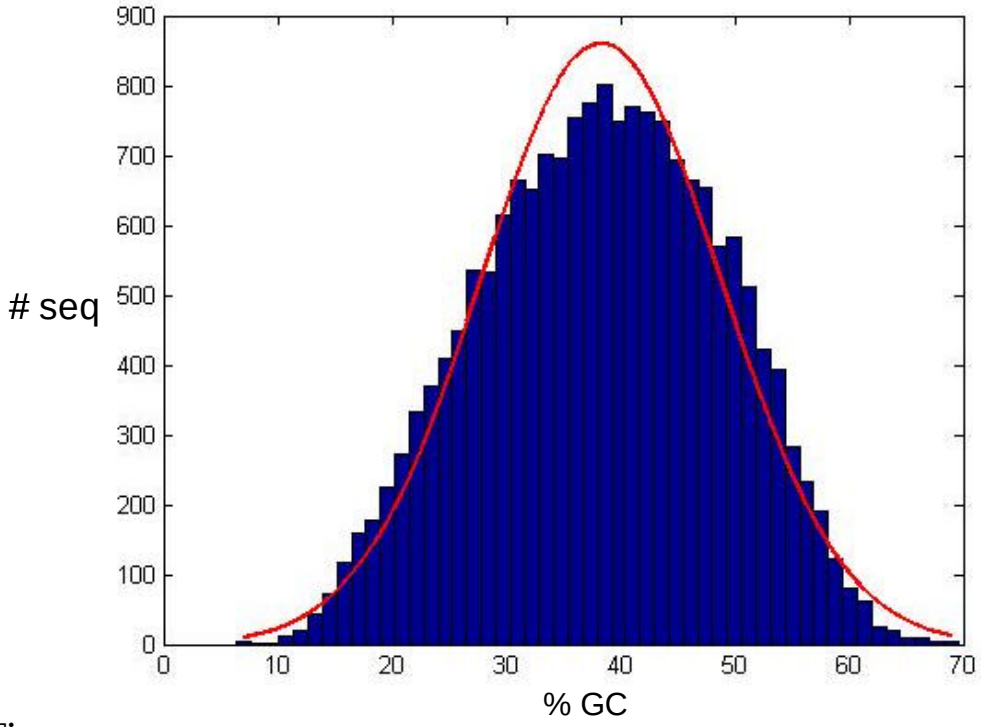


Figure A3. Distribution of GC content in mouse promoters. Red line represents fitted normal distributions (with mean 38.31 and standard deviation 10.52)

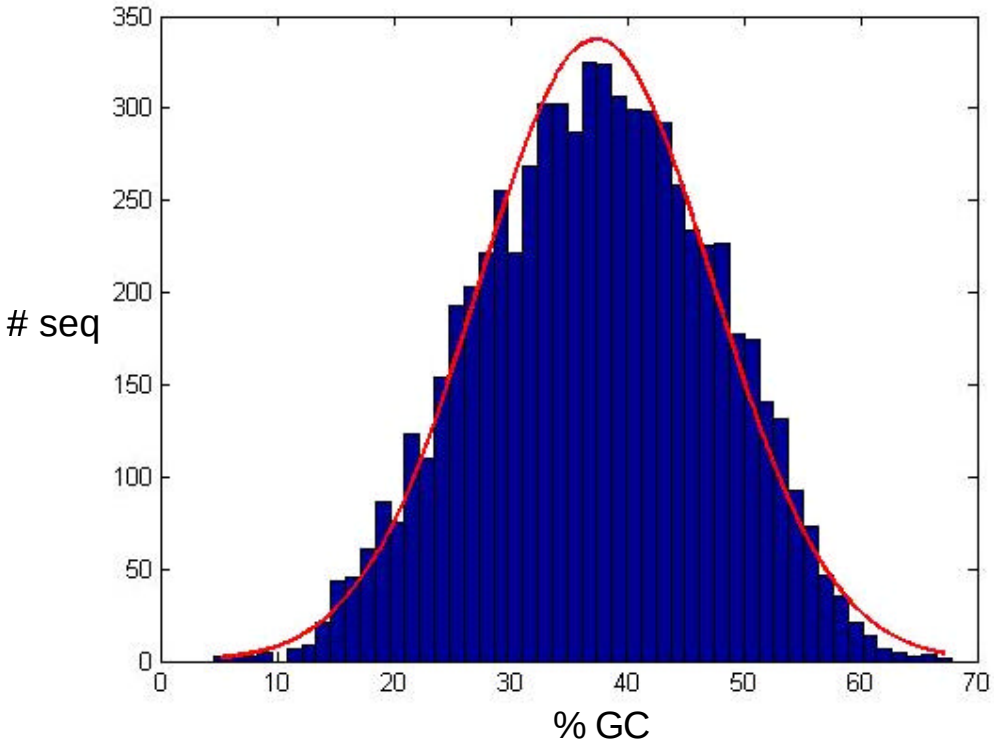


Figure A4. Distribution of GC content in rat promoters. Red line represents fitted normal distributions (with mean 37.38 and standard deviation 10.04).

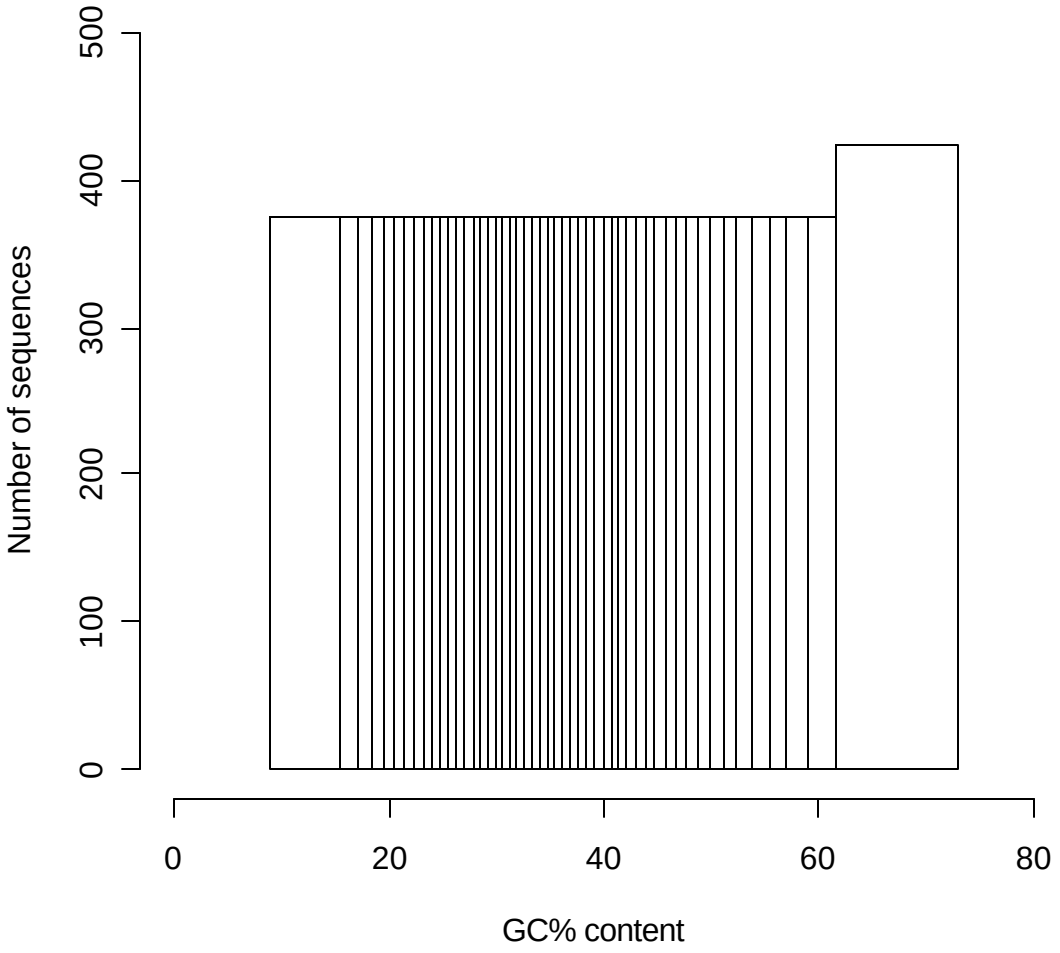


Figure A5. Distributions of GC content in human promoters, represented by a histogram of 50 bins each of which contains exactly 375 sequences except for the final bin that contains 424 sequences. The bin borders are given by the following array: 8.981424781, 5.48242334, 17.12593356, 18.41206602, 19.41747573, 20.44093631, 21.36474411, 22.29249012, 23.12616698, 23.98498806, 24.7291441, 25.47984645, 26.20016273, 27.00460829, 27.80918728, 28.46153846, 29.14967054, 29.84140234, 30.46171171, 31.15845539, 31.82561308, 32.55119454, 33.24845398, 33.94316855, 34.64765101, 35.33834586, 36.08159796, 36.78516229, 37.51434034, 38.29867675, 39, 39.85849057, 40.58823529, 41.41176471, 42.18403548, 43.12402698, 43.92567889, 44.80557168, 45.84569733, 46.84638861, 47.81105991, 48.85386819, 49.95495495, 51.23558484, 52.43632337, 53.92271663, 55.46651402, 57, 59.02891435, 61.63124641, 73.

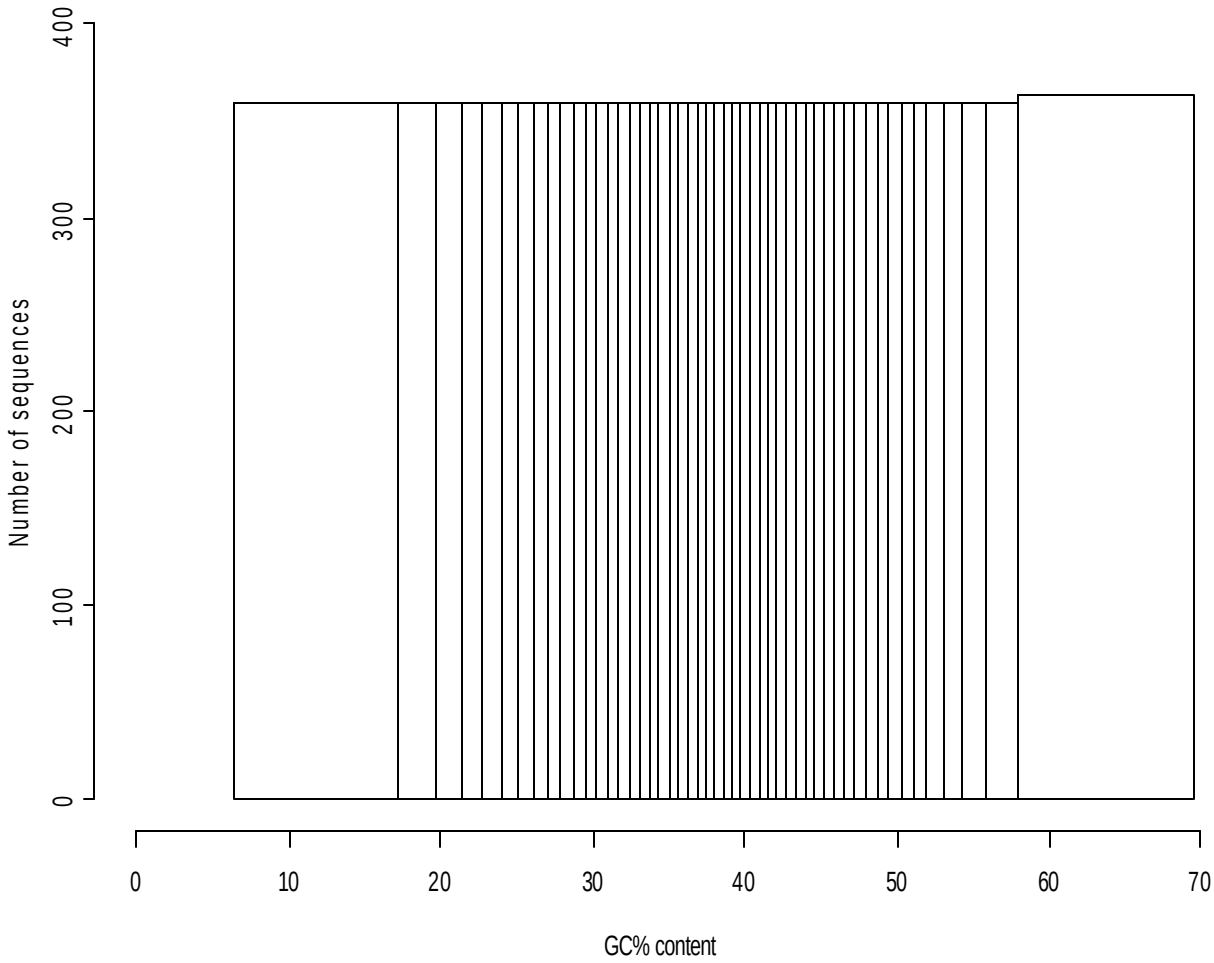


Figure A6. Distribution of GC content in mouse promoters, represented by a histogram made up of 50 bins each of which contains 359 promoters except for the final bin which contains 363. The bin borders are given by the following array: 6.352941176, 17.2403734, 19.63499056, 21.45005012, 22.78725459, 24.02251185, 25.10714286, 26.1328125, 27.07472775, 27.90522753, 28.77659574, 29.55508475, 30.2972561, 31.01289134, 31.6612141, 32.36404834, 33.04413429, 33.70245546, 34.35047951, 35.03348789, 35.6396217, 36.23529412, 36.83417085, 37.40388964, 38, 38.58823529, 39.13783324, 39.70588235, 40.34151547, 40.95908491, 41.55273438, 42.09115282, 42.70242393, 43.30949949, 43.92419175, 44.53870626, 45.18167457, 45.82352941, 46.50491046, 47.23360656, 47.92176039, 48.71520343, 49.47058824, 50.28735632, 51.17647059, 52, 53.11764706, 54.23529412, 55.76470588, 57.93103448, 69.55017301.

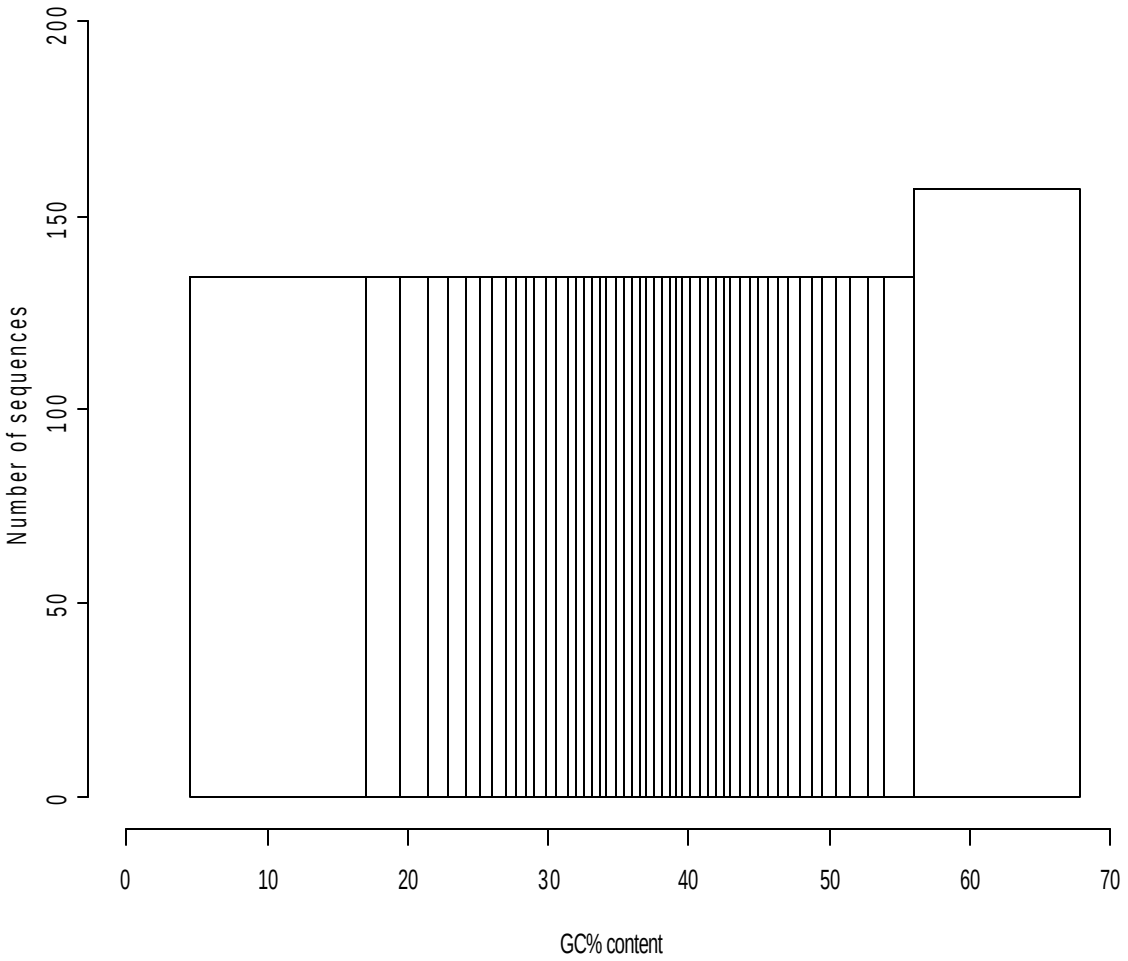
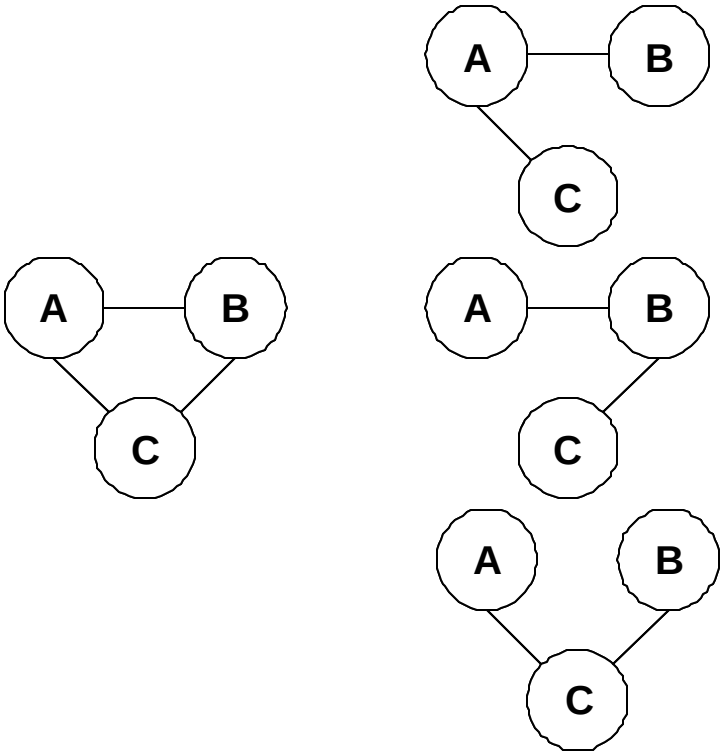


Figure A7. Distribution of GC content in rat promoters, represented by a histogram with a fixed number of 50 bins each of which contains 134 promoters except for the final bin which contains 157. The bin borders are given by the following array: 4.529411765, 16.98636469, 19.52821738, 21.3672391, 22.79411765, 24.12698413, 25.11578197, 25.98833441, 26.88277669, 27.65567766, 28.38883283, 29.07372401, 29.74322397, 30.60765191, 31.29147524, 31.89577718, 32.45967742, 33.05882353, 33.59240069, 34.12556054, 34.70588235, 35.3219697, 35.84715938, 36.42384106, 36.95968917, 37.49324689, 38.0010983, 38.55721393, 39.05411994, 39.59854015, 40.13909588, 40.74844075, 41.31355932, 41.94117647, 42.45951417, 43.01994302, 43.60097324, 44.26054458, 44.92273731, 45.60622914, 46.33885623, 47.05882353, 47.91785511, 48.66589327, 49.52941176, 50.50223214, 51.41176471, 52.65511459, 53.94117647, 55.92672414, 67.74916013.



i) fully connected graph represents full 3-order dependencies between transcription factors A,B and C

ii) Not fully connected graph represents partial 3-order dependencies between transcription factors A,B and C

Figure A8. Representation of higher order dependencies between transcription factors A, B and C.

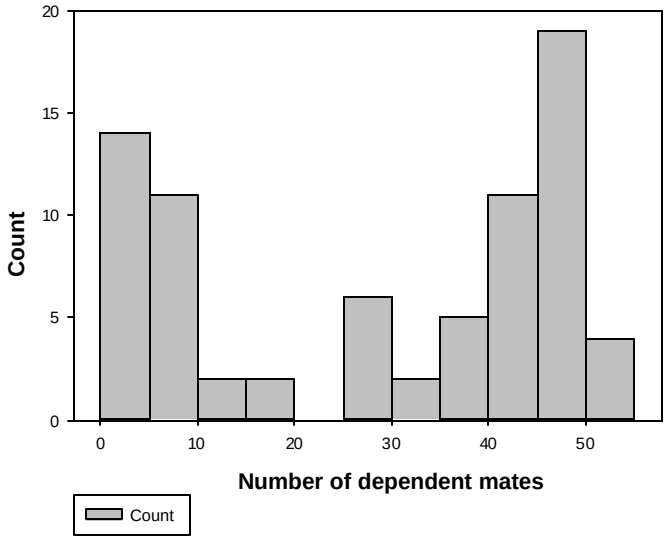


Figure A9. Histogram of number of dependent mates for each transcription factor in rat genome

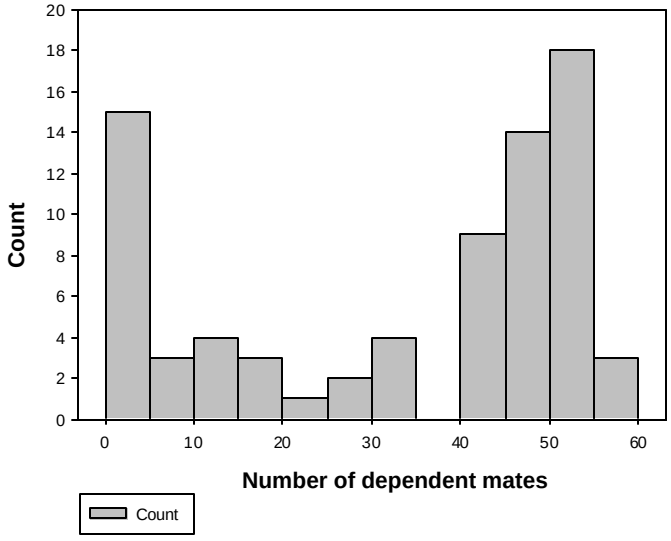


Figure A10. Histogram of number of dependent mates for each transcription factor in mouse genome

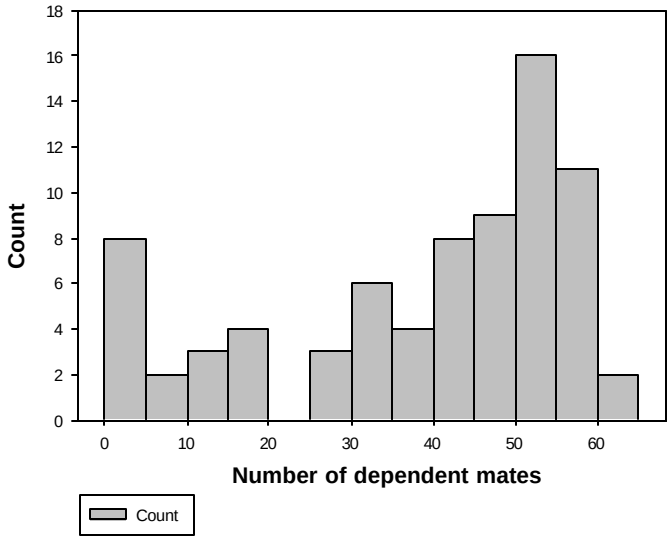


Figure A11. Histogram of number of dependent mates for each transcription factor in human genome

Table A1. The number of dependent mates for each transcription factor in human, mouse and rat genome, including cluster information (from [1]) about similarity between binding sites.

Name	TF-ID	# dep-mate-human	# dep-mate-mouse	# dep-mate-rat	Cluster (from [1])
Pparg	MA0066	0	0	0	Cluster-77
pax6	MA0069	0	0	0	Cluster-71
p53	MA0106	0	0	0	Cluster-78
Roaz	MA0116	0	0	0	-
pprag-rxra	MA0065	1	0	0	Cluster-91
hnf4a	MA0114	1	1	0	Cluster-32
nr1h2-rxra	MA0115	2	1	0	-
pax5	MA0014	3	2	0	Cluster-98
evi1	MA0029	8	6	0	Cluster-10
Srf	MA0083	6	5	1	Cluster-72
spi1	MA0080	39	0	2	Cluster-69
mzf1_1-4	MA0056	15	0	4	Cluster-85
myc-max	MA0059	16	10	4	Cluster-15
Nfya	MA0060	11	13	4	Cluster-46
gata2	MA0036	41	0	5	Cluster-9
TCF11-MafG	MA0089	33	7	5	Cluster-26
ddidt3-cebpa	MA0019	18	18	6	-

gata3	MA0037	31	0	7	Cluster-16
Spib	MA0081	14	18	7	Cluster-69
en1	MA0027	40	0	8	Cluster-8
ets1	MA0098	41	0	8	Cluster-2
zeb1	MA0103	13	13	8	-
pax2	MA0067	32	13	8	Cluster-30
yy1	MA0095	15	0	9	Cluster-94
nr3c1	MA0113	27	29	9	-
sox9	MA0077	28	18	12	Cluster-20
irf2	MA0051	32	34	13	Cluster-6
foxd3	MA0041	27	22	16	Cluster-17
Mafb	MA0117	31	26	17	-
sp1	MA0079	36	33	25	Cluster-74
mzf_1_513	MA0057	33	30	27	Cluster-99
bapx1	MA0122	47	41	28	-
rreb1	MA0073	43	44	28	Cluster-68
nkx2-5	MA0063	43	30	29	Cluster-41
esr1	MA0112	45	44	29	-
mef2a	MA0052	45	42	32	Cluster-82
T	MA0009	44	43	34	-
tlx1-nfic	MA0119	44	42	35	-
spz1	MA0111	42	42	36	-
rxra-vdr	MA0074	46	46	36	Cluster-91
nfkb1	MA0105	39	43	38	Cluster-5
Gfi	MA0038	48	47	38	Cluster-88
Sry	MA0084	52	40	40	Cluster-20
nhlh1	MA0048	38	47	41	-
nr2f1	MA0017	51	51	41	Cluster-2
hand1-tcfe2a	MA0092	48	47	42	Cluster-92
pbx1	MA0070	50	51	42	Cluster-73
elk1	MA0028	49	49	43	Cluster-2
nf-kappab	MA0061	45	48	44	Cluster-5
Mycn	MA0104	47	48	44	Cluster-15
nfil3	MA0025	51	49	44	-
Staf	MA0088	52	49	44	Cluster-81
runx1	MA0002	55	52	44	-
tal1-tcf3	MA0091	51	48	45	Cluster-7
foxd1	MA0031	50	49	45	Cluster-17
rora_2	MA0072	55	49	45	Cluster-93
Cebpa	MA0102	55	49	45	Cluster-14
Max	MA0058	50	50	45	Cluster-15
arnt-ahr	MA0006	50	51	45	Cluster-15
Rel	MA0101	51	50	46	Cluster-5

prrx2	MA0075	52	50	46	-
sox5	MA0087	55	45	47	Cluster-20
foxa2	MA0047	54	50	47	-
Gabpa	MA0062	54	52	47	-
Rela	MA0107	52	53	47	Cluster-5
Ar	MA0007	56	57	47	-
elk4	MA0076	52	52	48	Cluster-2
sox17	MA0078	54	52	48	-
foxq1	MA0040	55	52	48	-
usf1	MA0093	58	53	48	Cluster-15
tcf1	MA0046	58	54	48	Cluster-21
e2f1	MA0024	55	55	48	Cluster-35
rora_1	MA0071	59	52	50	Cluster-44
Hlf	MA0043	63	54	50	Cluster-23
Arnt	MA0004	58	55	51	Cluster-15
Creb	MA0018	61	52	53	Cluster-3

Table A2 - Scanning promoter sequences

General form of output after scanning promoter sequences for the given combination of transcription factors A and B.

transcription factors promoter sequence id (%GC)	**A**	**B**
Prom-id1 (P_{id1})%	A_1	B_1
Prom-id2 (P_{id2})%	A_2	B_2
….		
Prom-idn (P_{idn})%	A_n	B_n

Reference:

1. Kielbasa, S.M., D. Gonze, and H. Herzel, **Measuring similarities between transcription factor binding sites.** BMC Bioinformatics, 2005. 6: p. 237.

4.2 Computational prediction of transcription factor start sites

An additional practical application of results from descriptive data-mining about transcription factor site dependencies is that they can be used for *in silico* transcription start site prediction.

4.2.1 Results

In order to demonstrate this practical application, we selected three genes (ctmp, gap-43 and ngfrap) with well-characterized promoters in mouse (positive control; Genomatix GmbH, Munich, Germany) but not in rat. In addition, the promoters of these rat genes did not belong to the set of upstream regions which we used to detect dependencies. We used 2,000 bp upstream, and ~100 bp downstream, of the annotated (by Genomatix) mouse start and ~6,000 bp of 5' sequence upstream of exon one of the rat gene ortholog. We identified the best matching window between the mouse and rat modules, and used the relative distance between conserved transcription factor patterns and the start of transcription in mouse to predict the corresponding start in rat (Table 1). Then we experimentally determined the actual starts for both the mouse and rat genes using 5'-RACE on total brain RNA by sequencing (Figure 1 and 2, Table 1).

The third column of Table 1 shows the predicted position (ppos) of transcription initiation, and the absolute error of this prediction:

$$error = |pos - ppos| \quad (8)$$

Table 1. Experimentally verified and computational predictions of the transcription start site on rat genes ctmp, gap-43 and ngfrap.

Gene name chromosome/strand	**Experimentally verified position of transcription initiation (pos)**	**Predicted start of transcription start site (ppos)**	**absolute error |pos-ppos|**
ctmp chr 2 / "+"	pos = 189296542	ppos = 189296401	141
gap-43 chr 11 / "+"	pos = 59989243	ppos = 59989221	22
ngfrap chr X / "+"	pos = 123586283	ppos = 123586556	273

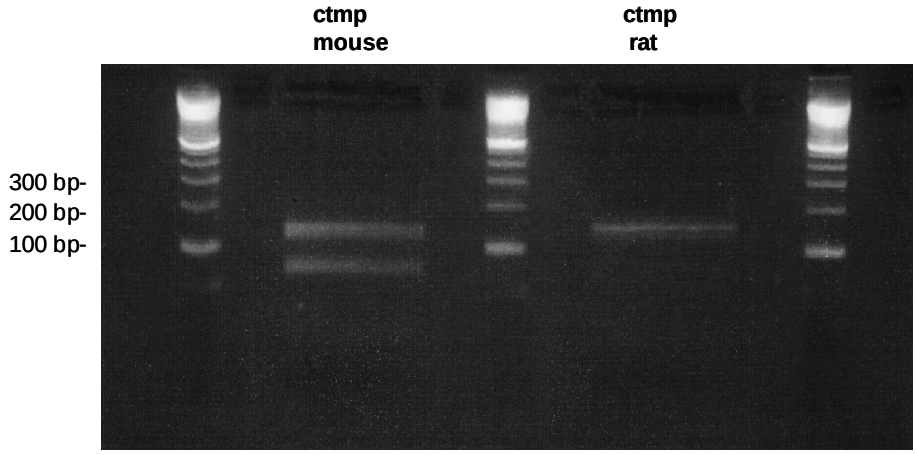


Figure 1. PCR gel for the mouse and rat 5' end of gene ctmp. Mouse was used as a positive control, since the information about TSS is known.

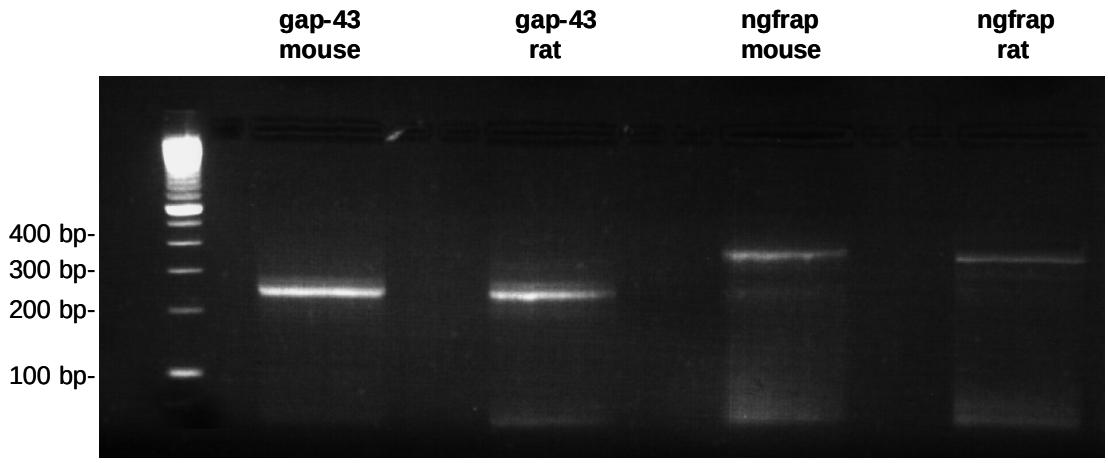


Figure 2. PCR gel for the mouse and rat 5' end of genes gap-43 and ngfrap. Mouse was used as a positive control, since the information about its TSS was known.

We investigated the sequence similarity between each of the mouse and rat promoter sequences around the transcription start sites. We took 400bp upstream of the start and ran the blast 2 sequence program on them [1]. There was no significant alignment between the mouse and rat versions of either ctmp or ngfrap. There was a short (41/42 bp) homology between the mouse and rat gap-43 promoters. These results showed that there was insufficient similarity between the promoters to identify the starts by sequence homology alone (or it can be done but error larger then 400bp). All other available tools for predicting the transcription start site [2-5] use sequence-based algorithms (without using any information about the transcription start site of ortholog genes). Accordingly, the comparison of the results obtained is not completely fair (Table 2). However, with this approach we demonstrate a new computational strategy for predicting transcription initiation. In addition, the idea of this application was to demonstrate the usefulness of the information obtained about dependencies in the computational promoter analysis, confirmed experimentally.

Table 2. Prediction of start sites (best predictions, January 2008) for the rat genes: ctmp, gap-43,ngfrap by Promoter 2.0 [2], Dragon Promoter Finder [3], WWW Promoter Scan [4] and NNPP [5].

Gene	Predicted start (and absolute error) of transcription by			
	Promoter 2.0	**Dragon PromoterFinder**	**WWW Promoter Scan**	**NNPP**
ctmp	Ppos= 89292824 error =3718	ppos=no prediction error = 8	ppos= 189295908 error =634	ppos= 189294725 error=1817
gap-43	Ppos = 59987984 error =1259	ppos=no prediction error=8	ppos= no prediction error =8	ppos=59985353 error=3890
ngfrap	ppos=123580105 error =6178	ppos=no prediction error = 8	ppos= 123586024 error =259	ppos=123584655 error= 1628

4.2.2 Methods

Computational prediction of transcription start sites

For the computational prediction of transcription start sites we employed a combination of comparative genomics and the scanning computational prediction of groups of binding sites of dependent transcription factors. If we have two orthologous genes where we know the transcription start site for one but not for the other, we can first try a comparative genomics approach through sequence comparison in order to estimate where the transcription start site is. However, very often this is impossible in practice, because unambiguous alignment of sequences showing very little conservation can be difficult. An alternative is the following strategy: for the known promoter (Seq1, length ~1-2Kbp), predict all cis-regulatory modules. Then, for the gene for which we do not know the transcription start site, take ~6-8Kbp of sequence 5' of exon one (Seq2) (with the condition that Seq2 is longer than Seq1). Predict binding sites of all 2-order dependent transcription factors in Seq1 and Seq2 and find the best-matched window with a length less than or equal to Seq1 in sequence Seq2. The best-matched window contains the

highest number of the same transcription factor groups conserved between Seq1 and Seq2. In order to describe the best-matched window, we introduce a scoring function F:

$$F(Seq1, SSeq_i 2) = \frac{c(SSeq_i 2)}{c(Seq1)} \qquad (1)$$

where SSeqi2 is the subsequence of sequence Seq2 with the length of Seq1 (2kbp) and start i (1=i=length(Seq2)-length(Seq1)), c(Seq1) is the number of gorups in promoter Seq1, and c(SSeqi2) is the number of conserved groups in sequence SSeqi2. The scoring function has a value between 0 and 1. The window (subsequence of sequence Seq2), for which the scoring function is highest corresponds to the best-matched window, i.e.:

$$BMSeq2 = \underset{i}{\operatorname{argmax}}\, F(Seq1, SSeq_i 2) \qquad (2)$$

Using the relative distance between average start binding sites of conserved transcription factor patterns, and the start of transcription in sequence Seq1, the corresponding start in BMSeq2 can be predicted, i.e. the start in sequence Seq2 (Figure 3).

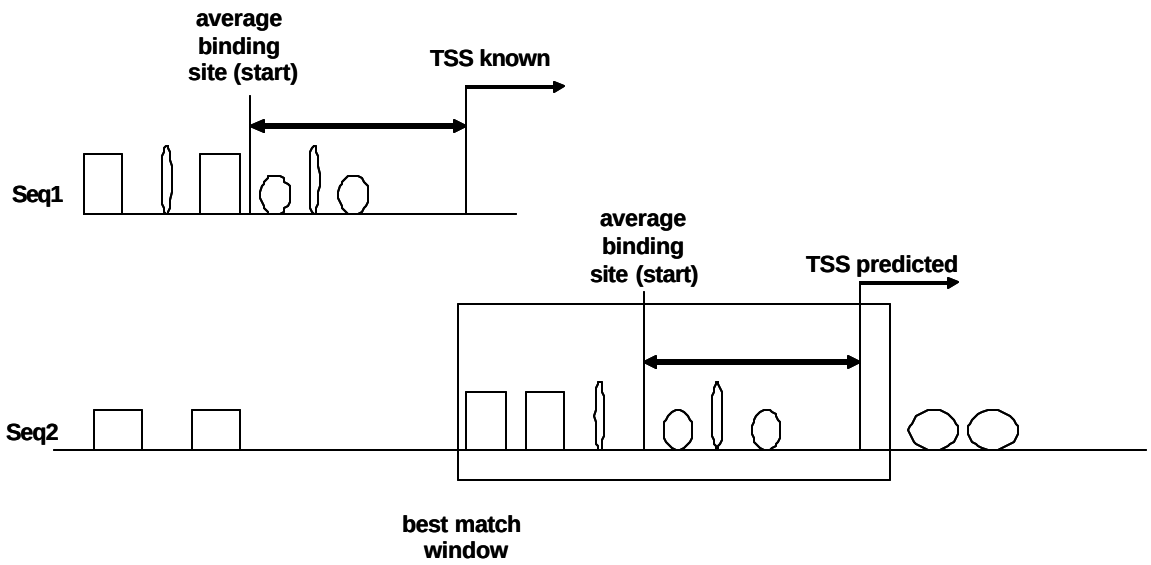


Figure 3. Prediction of the transcription start site (TSS) on Seq2 based on the conservation of predicted cis-regulatory modules with Seq1, for which the TSS is known.

Wet-lab identification of transcription start site

Experimental determination of the transcription start site for both the mouse and rat genes used in this work was performed by 5'-RACE on total brain RNA, followed by sequencing. RNA isolation from mouse and rat brain was performed using Trizol reagent (Invitrogen AG, Basel, Switzerland) according to the manufacturer's instructions. 5'-RACE was performed using the SMARTTM RACE cDNA Amplification kit from Takara Bio Company (Mountain View, USA) according to the manufacturer's instructions. Mouse genes were a positive control in this experiment to confirm the Genomatix starts. The following gene-specific primers (Microsynth AG –Balgah, Switzerland) were used for the 5'RACE: gap-43 (mouse and rat) GCAACGGGAGCACATCCTTCTCCTT; ngfrap (mouse and rat) TTCTCCGGATCTCTCTCATCTCCTCCA; ctmp (mouse and rat) CCAGCTGGGGTTAGGGAGAGCATAGTCC. All PCR bands were gel-purified (QIAquick Gel Extraction Kit, QIAGEN AG, Hombrechtikon, Switzerland) and sequenced either using our in-house facility or else Microsynth AG, Balgah, Switzerland.

References

1. Tatusova, T.A. and T.L. Madden, **BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences.** *FEMS Microbiol Lett*, 1999. **174**(2): p. 247-50.
2. Knudsen, S., **Promoter2.0: for the recognition of PolII promoter sequences.** *Bioinformatics*, 1999. **15**(5): p. 356-61.
3. Bajic VB, Seah SH, Chong A, Zhang G, Koh JL, Brusic V: **Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters**. *Bioinformatics* 2002, **18**(1):198-199.
4. Prestridge, D.S., **Predicting Pol II promoter sequences using transcription factor binding sites**. *J Mol Biol*, 1995. **249**(5): p. 923-32.

5. Reese, M.G., **Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome**. Comput Chem, 2001. **26**(1): p. 51-6.

5. Conclusions and Perspectives

The first main finding reported in this thesis is that it is wrong to assume, *a priori*, that positions in transcription factor binding sites are all either independent or dependent of one another. Position dependencies should be tested using rigorous statistical methods on a case-by-case basis. When dependencies are detected, they can be modelled in a very simple way, which doesn't require complex mathematical tools with a lot of parameters and more data. An example of such a model, including a web-based implementation of the algorithm, is reported in this thesis (http://promoterplot.fmi.ch/cgi-bin/dep.html). A possible biological explanation of position dependencies in transcription factor binding sites is given by exploring the 3D structure of DNA-protein complexes. It has been shown that the conformational energy (indirect readout) of DNA in complexes with transcription factors which have dependent positions in their binding sites is significantly higher than in those with transcription factors which do not have dependent positions in their binding sites. How to use modelling dependencies for the scanning method for the computational prediction of transcription factor binding sites has also been demonstrated. However, the proposed method can also be integrated into *ab initio* methods for transcription factor binding sites. The biggest advantage of *ab initio* methods is that they do not require any database of known binding sites. The next step will be the implementation/integration of the method for modelling dependencies into tools for the *ab initio* prediction of transcription factors based on over-represented motifs in the input sequences (MEME, Gibs sampling). In addition, computational detections of other signals in RNA/DNA sequences (such as splicing regulatory elements and translation regulatory elements) can also be based on a similar methodology, described in this thesis, with small modifications and adaptations for each specific problem.

The structural analysis of DNA-protein (with one or more proteins) and protein-protein complexes confirmed that protein-protein and protein-DNA interface parameters, such as the interface area and the number of interface residues/atoms and hydrogen bonds, and the distribution of interface residues, hydrogen bonds, van der Waals contacts and secondary structure motifs in complexes where multiple proteins are bound to DNA are

no different in protein-protein, single protein-DNA or multiple protein-DNA complexes. Thus, if we have two (or more) proteins which bind together, there will be no influence on these interface parameters. Also, if we have one protein bound to DNA, then that binding will have no influence (in terms of the interface parameters mentioned) on the types of interface interactions that can occur with subsequent protein-protein complex expansion. Distortion is significantly higher when multiple proteins bind to DNA. This distortion is required to accommodate multiple protein-binding events. The combinatorial assembly of transcription factors has been known for a long time to play an important role in stabilising regulatory complexes. A deeper understanding of structural considerations may be helpful when predicting the assembly of transcription factor complexes. The formation of multiple protein interactions with DNA results in a decrease in protein-protein affinity and an increase in protein-DNA affinity, with a net gain in overall stability for a protein-protein-DNA complex. Such effects are clearly important for modelling transcription factor cooperativity.

In addition, the physical overlap of two factors does not simply relate to the region on the DNA where the binding site is found. Two factors may lie very close together but not physically overlap because their side-chains can interlink with one another. In this way, it is possible to find a large overlap between two transcription factor binding sites, but from a 3D perspective it is still possible for both factors to bind simultaneously. It may also be that one transcription factor binds to the minor and another to the major groove of DNA. That information is also useful for modelling transcription factor cooperativity. It has been confirmed that structural data can be useful for *in silico* promoter analysis. All available information should be integrated in order to make more realistic models for the simulation of a biological process.

Next, this thesis reports the results from a computational prediction of dependencies between transcription factors which usually act together in gene regulation in the human, mouse and rat genomes. Constructing higher order dependencies/cooperativites between transcription factors (network) in human, mouse and rat is very important for understanding the gene expression process (http://promoterplot.fmi.ch/TFDEP1/). The strategy for this prediction is based on the scanning method which incorporates position dependencies (described in chapter 2 of the thesis) plus some structural information (as

indicated above and in chapter 3). It has been shown in laboratory research that many transcription factors interact together in the regulation of gene expression. These discoveries have been used as positive controls for the computational prediction of transcription factor dependencies reported in this thesis. In combination with data from genomics, proteomics and matabolomics projects, the proposed network can be expanded. It is shown that modelling transcription factor cooperativities (dependencies) improves the quality of transcription factor binding sites (http://promoterplot.fmi.ch/TFDepSSeq1/). In addition, the transcription start site can be predicted computationally using comparative genomics and groups of binding sites of dependent transcription factors. This thesis reports examples of such *in silico* TSS predictions for three genes (ctmp1, gap43 and ngfrap) expressed in mouse/rat brain, including laboratory validation.

Finally, the Bayesian method for the detection of dependencies between positions in transcription factor binding sites can easily be converted into a method for estimating the quality of multiple sequence alignments. That method is simple, with linear complexity, which is easy to implement and which performs better than other state-of-the-art methods which are more complex. A web-based implementation of this method for the quality estimation of multiple sequence alignments is freely available from http://www.fmi.ch/groups/functional.genomics/tool.htm. It can be integrated into any tool that uses statistical estimates of sequence alignments or as a post-processing filter of the output from any tool that returns a number of ordered alignments. Possible applications include: motif finder algorithms; algorithms for profile-profile and sequence-profile alignment; and the analysis of protein domains and their families. This gives space for further work which can be based on the method proposed in this thesis.

Computational analysis of promoter activity and DNA-protein interactions is useful for understanding many crucial cellular processes, including transcription, recombination and replication. Some of the techniques for the computational analysis of promoter and DNA-protein interactions are predominantly useful for data-mining the huge amount of data produced in the laboratory. In that way, the computational analysis of promoters and DNA-protein interactions helps towards additional understanding and analysis of data

produced in the laboratory (describing the data, and determining some rules, associations and patterns). On the other hand, computational methodology for the analysis of promoters and DNA-protein interactions can assist laboratory work before it has been performed. Different kinds of computational predictions and simulations can be beneficial for designing optimal and effective laboratory work. In this thesis, both aspects of the computational analysis of promoters and DNA-protein interactions are covered. In addition, computational approaches for analysing promoters and DNA-protein interactions will become more powerful as more and more complete genome sequences, 3D structural data and different kinds of high-throughput data become available. Modern scientific research into promoters and DNA-protein interactions represents a high level of co-operation between computational and laboratorial methods. This union will be even stronger in future research.

Research in bioinformatics and computational biology is important from the practical aspect (biological point of view) in order to simulate, explain, analyse and predict different cellular processes, but on the other hand it assists us in the development of methodologies and algorithms in mathematics and computer science. For some biological problems, computational methodologies have not yet been developed; for some of them there are computational tools which can be applied but which require additional adaptations and specifications of the given tool. In this thesis I have attempted to cover both the theoretical and practical questions by bridging the gap from theoretical methodological research to defined biological applications.

APPENDIX A

PAPER II – Quality estimation of multiple sequence alignments by Bayesian hypothesis testing

This chapter of the thesis reports how one statistical method for the analysis of transcription factor binding sites can be used for estimating the quality of multiple sequence alignments. In particular, the Bayesian method for the detection of dependencies between positions in transcription factor binding sites can easily be converted into a method for estimating the quality of multiple sequence alignments. That method is simple, of linear complexity, easy to implement and performs better than other state-of-the-art methods which are more complex.

All supplementary materials from this paper are given in this chapter, and implementation of the method is freely available from http://www.fmi.ch/groups/functional.genomics/tool.htm

BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 18 2007, pages 2488–2490
doi:10.1093/bioinformatics/btm366

Sequence analysis

Quality estimation of multiple sequence alignments by Bayesian hypothesis testing

Andrija Tomovic and Edward J. Oakeley*

Friedrich Miescher Institute for Biomedical Research, Novartis Research Foundation, Maulbeerestrasse 66, CH-4056 Basel

Received on May 17, 2007; revised on July 2, 2007; accepted on July 9, 2007

Advance Access publication July 27, 2007

Associate Editor: Alex Bateman

ABSTRACT

Summary: In this work we present a web-based tool for estimating multiple alignment quality using Bayesian hypothesis testing. The proposed method is very simple, easily implemented and not time consuming with a linear complexity. We evaluated method against a series of different alignments (a set of random and biologically derived alignments) and compared the results with tools based on classical statistical methods (such as sFFT and csFFT). Taking correlation coefficient as an objective criterion of the true quality, we found that Bayesian hypothesis testing performed better on average than the classical methods we tested. This approach may be used independently or as a component of any tool in computational biology which is based on the statistical estimation of alignment quality.

Availability: http://www.fmi.ch/groups/functional.genomics/tool.htm

Contact: edward.oakeley@fmi.ch

Supplementary information: Supplementary data are available from http://www.fmi.ch/groups/functional.genomics/tool-Supp.htm

1 INTRODUCTION

Statistical estimation of the significance of proposed alignments is one of the central challenges of evaluating the output of all alignment tools. Local ungapped alignments play an important role in the discovery and classification of both DNA and protein sequences. To evaluate a proposed sequence alignment we must know the likelihood of it occurring by chance rather than, for example, deriving from a common ancestral sequence. Statistically significant alignments have a higher chance of being biologically relevant. The evaluation of ungapped local alignment is usually made using its information content or relative entropy (Hertz and Stormo, 1999; Nagarajan *et al.*, 2005):

$$I_{\text{seq}} = \sum_{i=1}^{L} \sum_{j=1}^{|A|} \frac{n_{ij}}{n} \log \frac{n_{ij}/n}{b_j} \tag{1}$$

where L is the length of the sequence from an alphabet A, n_{ij} count of the j-th letter in the i-th column of alignment, n is the number of sequences in the alignment and b_j the background frequency of the j-th letter. Using this scoring function (1) and a null model, which assumes that each of the k columns has n letters independently sampled according to the background distribution we can estimate a P-value. The P-value for a given scoring value s_0 represents the probability of an entropy score of s_0 or better under the null model (Hertz and Stormo, 1999; Nagarajan *et al.*, 2005). When the information content (I_{seq}) is small and the number of sequences (n) is large, the value $2nI_{\text{seq}}$ tends to be χ^2-distributed with $k(|A|\text{-}1)$ degrees of freedom (Wilks, 1938). But this approximation is very poor when we have large scores and few sequences, which is a common situation. Several methods have been developed to improve this P-value estimation (Dembo *et al.*, 1994; Hertz and Stormo, 1999; Karlin and Altschul, 1990; Keich, 2005; Nagarajan *et al.*, 2005). In this work, we present a web-based tool for estimating sequence alignment significances without gaps using Bayesian hypothesis testing. Bayesian methods have already been used in algorithms for sequence alignment (Liu and Lawrence, 1999; Liu *et al.*, 1995; Lunter *et al.*, 2005; Suchard and Redelings, 2006; Webb *et al.*, 2002; Zhu *et al.*, 1998), but in our implementation we used a Bayesian approach to evaluate multiple sequence alignments without gaps that had already been generated. This approach can be used independently or as a component of any tool in computational biology which uses statistical alignment quality estimates.

2 METHOD

Quality estimation of multiple sequence alignments by Bayesian hypothesis testing is based on the work (Minka, 1998; Liu and Lawrence, 1999) which we have adapted for use with DNA and protein sequence alignments. In the interest of simplicity, we will demonstrate the utility of this method in the context of DNA sequence alignments, but it can easily be applied to protein sequence alignments too.

Let us define an alignment X of n DNA sequences of length L:

$$\begin{matrix} X_1^1 X_2^2 \ \ldots \ X_L^1 \\ X_1^2 X_2^2 \ \ldots \ X_L^2 \\ \ldots \\ X_1^n X_2^n \ \ldots \ X_L^n \end{matrix} \tag{2}$$

Let X_i represent the vector frequencies for each letter (base) for column i of a multiple alignment: $X_i = [X(a,i), X(c,i), X(g,i), X(t,i)]$. We also define Y as a vector with the same length as X_i, $Y = [Y(a), Y(c), Y(g),$

*To whom correspondence should be addressed.

Table 1. Summary of results from the estimation of 207 alignments (100 random and 107 JASPAR-derived) produced by three methods sFFT, csFFT and Bayes method

Method	True positive	True negative	False positive	False negative	Specificity	Sensitivity	Corr. coef.
sFFT	107	60	40	0	0.60	1	0.66
csFFT	107	60	40	0	0.60	1	0.66
Bayes method	97	100	0	10	1	0.91	0.91

$Y(t)] = [y_a n,\ y_c n,\ y_g n,\ y_t n]$, where y_a, y_c, y_g and y_t represents the background frequencies of each base, respectively a, c, g, and t. Background frequencies of each base can be estimated based on input data or user can specify it. To evaluate the alignment (2), first we will test the following hypotheses:

$$\begin{aligned} &\mathrm{H}_0\text{: } Y \text{ and } X_i \text{ come from the same multinomial distribution} \\ &\mathrm{H}_1\text{: } Y \text{ and } X_i \text{ come from different multinomial distributions} \end{aligned} \tag{3}$$

This hypothesis testing can be evaluated directly [in a way similar to that described by (Liu and Lawrence, 1999; Minka, 1998), or in the form of a test for independence (Minka, 1998) which gives slightly different results because of different priors. We have used second approach and a detailed description as to how it is possible to convert the hypothesis test (3) into an independence test is given in Supplementary Material 1. For each column we calculated a Bayes factor $BF_i(\mathrm{H_o};\ \mathrm{H_1})$ and likelihoods $P_i(Y,X_i|\mathrm{H_0})$ and $P_i(Y,X_i|\mathrm{H_1})$. Because of our assumption of independence between the columns, after calculating $BF_i(\mathrm{H_o};\ \mathrm{H_1})$ and $P_i(Y,X_i|\mathrm{H_0})$ and $P_i(Y,X_i|\mathrm{H_1})$ for each $i = 1, \ldots, L$ (for each column) we can calculate:

$$\mathrm{BF} = \prod_{i=1}^{L} \mathrm{BF}_i(\mathrm{H_0},\mathrm{H_1}) \tag{4}$$

$$P(\mathrm{H_0}|Y,X) = \frac{P(\mathrm{H_0}) \prod_{i=1}^{L} P_i(Y,X_i|\mathrm{H_0})}{P(\mathrm{H_0}) \prod_{i=1}^{L} P_i(Y,X_i|\mathrm{H_0}) + P(\mathrm{H_1}) \prod_{i=1}^{L} P_i(Y,X_i|\mathrm{H_1})} \tag{5}$$

These scores provide us with an estimate of the multiple sequence alignment significance. It is more significant when BF is small (much smaller than 1) and when the posterior probability of the null model $P(\mathrm{H_o}|\ Y,\ X)$ is small (smaller probability of null model for the given alignment, i.e. smaller probability that given alignment is random). Jeffreys' scale (Jeffreys, 1961) of evidence for Bayes factors is given in Supplementary Material 1- Table 2. We used the posterior probability of the random model (null hypothesis) as a final score of alignment quality for the evaluation of our method (see the next section), because it is a more precise score value than Bayes factor.

3 RESULTS AND DISCUSSION

In this section we report our evaluation of the presented method and its comparison to other methods from classical (orthodox) statistics. We took 107 alignments of transcription factor binding sites, representing each factor in the JASPAR database (Lenhard and Wasserman, 2002; Sandelin *et al.*, 2004) and calculated the BF (4) and posterior probability of the null hypothesis (random model) (5). Detailed list for each transcription factor and its corresponding posterior probability and Bayes factor is given in Supplementary Material 2. All alignments, but 10, were found to be significant with very small posterior probabilities for the null hypothesis (much smaller than 0.001). Next, we generated 100 random alignments (available from http://www.fmi.ch/groups/functional.genomics/RandomAlignments.zip) using the RSA tool (van Helden, 2003). The random and JASPAR alignments had approximately the same distribution in terms of length and the number of sequences (Supplementary Material 3-Table 1). For each random alignment, we calculated BF (4) and posterior probabilities of the null hypothesis (5) (Supplementary Material 4). All alignments had posterior probabilities higher than 0.99 and they are correctly identified as not being statistically significant (true negatives). There are several classical (orthodox) techniques for the statistical evaluation of local ungapped alignments. Fast, but inaccurate, techniques are used in motif discovery tools [e.g. MEME (Bailey and Elkan, 1994), Consensus (Hertz *et al.*, 1990; Hertz and Stormo, 1999)]. In Supplementary Material 5 - Table 1, we report some of the more accurate methods for the statistical estimation of short ungapped alignments and their running times. The time complexity for the calculation of Bayes factor (4) and posterior probability (5) is linear O(L), and this has advantages over these other methods. We compared results (posterior probabilities of the random model) obtained by Bayesian approach with the *P*-values calculated by two classical methods csFFT (Nagarajan *et al.*, 2005) and sFFT (Keich, 2005) for a the transcription factor binding site alignments of each factor in the JASPAR database and 100 random alignments. In Table 1 we summarize the results for 207 alignments based on the *P*-values provided by the sFFT and csFFT methods, together with the results from the Bayesian method. The calculation of specificity and sensitivity was performed using the following formula:

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \quad \mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

Finally, Pearson product-moment correlation coefficients [also called the 'phi coefficient of correlation' (Burset and Guigo, 1996; Tompa *et al.*, 2005)] were calculated using:

$$\mathrm{Corr.Coef.} = \frac{\mathrm{TP}*\mathrm{TN} - \mathrm{FN}*\mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})*(\mathrm{TN} + \mathrm{FP})*(\mathrm{TP} + \mathrm{FP})*(\mathrm{TN} + \mathrm{FN})}} \tag{7}$$

Correlation coefficients may take any value between -1 (indicating perfect anticorrleation) and 1 (indicating perfect correlation).

We conclude, based on Table 1, that the Bayesian approach is superior to the classical approaches.

4 CONCLUSIONS

The method for using Bayesian hypothesis tests to evaluate alignment quality is simple, easy to implement and has a linear time complexity. Our method shows very high sensitivity and specificity in distinguishing biologically relevant from random alignments. It performs much better than methods based on classical statistics (Table 1). It can be integrated into any tool that uses statistical estimates of sequence alignments or as a post-processing filter of the output from any tool that returns a number of ordered alignments. Possible applications include: motif finder algorithms; algorithms for profile–profile and sequence–profile alignment; and the analysis of protein domains and their families. Our tool is available at http://www.fmi.ch/groups/functional.genomics/tool.htm.

ACKNOWLEDGEMENTS

We would like to thank Michael Stadler for helpful discussions. This work was supported by the Novartis Research Foundation. Funding to pay the Open Access publication charges was provided by the Novartis Research Foundation FMI.

Conflict of interest: none declared.

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Burset,M. and Guigo,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.

Dembo,A. *et al.* (1994) Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Prob.*, **22**, 2022–2039.

Hertz,G.Z. *et al.* (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.

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

Jeffreys,H. (1961) *Theory of Probability*. Clarendon Press, Oxford.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Keich,U. (2005) sFFT: a faster accurate computation of the p-value of the entropy score. *J. Comput. Biol.*, **12**, 416–430.

Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.

Liu,J.S. and Lawrence,C.E. (1999) Bayesian inference on biopolymer models. *Bioinformatics*, **15**, 38–52.

Liu,J.S. *et al.* (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Association*, **90**, 1156–1170.

Lunter,G. *et al.* (2005) Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, **6**, 83–93.

Minka,T. (2003) Bayesian inference, entropy, and the multinomial distribution. *Technical Report*.

Nagarajan,N. *et al.* (2005) Computing the *P*-value of the information content from an alignment of multiple sequences. *Bioinformatics*, **21** (Suppl. 1), i311–i318.

Sandelin,A. *et al.* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

Suchard,M.A. and Redelings,B.D. (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, **22**, 2047–2048.

Tompa,M. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.

Webb,B.J. *et al.* (2002) BALSA: Bayesian algorithm for local sequence alignment. *Nucleic Acids Res.*, **30**, 1268–1277.

Wilks,S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, **9**, 60–62.

Zhu,J. *et al.* (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, **14**, 25–39.

Supplementary material 1

for the paper "Quality estimation of multiple sequence alignments by Bayesian hypothesis testing"

A. Tomovic and E. J. Oakeley

The hypothesis test (3 in the paper) can also be converted into a test for independence in the following way. We introduce a random variable Q_i which takes values {1, 2}. When Q_i=1 it represents the distribution X_i and when Q_i=2 it represents Y. A second variable B_i takes its values from the DNA sequence alphabet {a, c, g, t} as shown in Table 1.

Table 1. Relationships between random variables B_i and Q_i

B_i / Q_i	B_i=a	B_i=c	B_i=g	B_i=t
Q_i=1	X(a,i)	X(c,i)	X(g,i)	X(t,i)
Q_i=2	Y(a)	Y(c)	Y(g)	Y(t)

From this, we can test the following hypotheses (equivalent to (3)):

H_0: Q_i and B_i are independent random variables
H_1: otherwise

When Q_i and B_i are independent, X_i and Y will not share the same distribution. For Bayesian hypothesis testing we can calculate the Bayes factor $BF_i(H_o; H_1)$ as follows:

$$BF_i(H_0; H_1) = \frac{P_i(B_i, Q_i \mid H_0) P_i(H_0)}{P_i(B_i, Q_i \mid H_1) P_i(H_1)} \quad (1)$$

Assuming that the *a priori* probabilities for the models (hypotheses) H_0 and H_1 are equal ($P_i(H_0) = P_i(H_1) = 0.5$), the Bayes factor will be:

$$BF_i(H_0;H_1) = \frac{P_i(B_i, Q_i \mid H_0)}{P_i(B_i, Q_i \mid H_1)} \tag{2}$$

Because B_i and Q_i are independent under the null hypothesis, the Bayes factor will be:

$$BF_i(H_0;H_1) = \frac{P_i(B_i \mid H_0)P_i(Q_i \mid H_0)}{P_i(B_i, Q_i \mid H_1)} \tag{3}$$

The index i implies that this Bayes factor is calculated for column i. Then, we can use the fact that:

$$P_i(B_i \mid H_0) = \int_{\bar{p}} P(B_i, \bar{p} \mid H_0) \tag{4}$$

where $\bar{p}$ is a vector of [P(a),P(c), P(g), P(t)].

A conjugate prior for $\bar{p}$ is the Dirichlet distribution (5):

$$P(\bar{p} \mid a) \sim Dir(a_a, a_c, a_g, a_t) = \frac{\Gamma(\sum_b a_b)}{\prod_b \Gamma(a_b)} \prod_b P(b)^{a_b - 1} \tag{5}$$

where P(b)>0 and $\sum_b P(b) = 1$. Given a Dirichlet prior, the joint distribution of B_i and $\bar{p}$ is:

$$P(B_i, \bar{p} \mid a) = \frac{\Gamma(\sum_b a_b)}{\prod_b \Gamma(a_b)} \prod_b P(b)^{B_i(b) + a_b - 1} \tag{6}$$

where $B_i(b) = X(b,i) + Y(b)$, i.e. the column sum.

And the posterior is:

$$P(\bar{p} \mid B_i, a) \sim Dir(B_i(b) + a_b) \tag{7}$$

And finally we can calculate

$$P_i(B_i \mid H_0) = \int_{\bar{p}} P(B_i, \bar{p} \mid H_0) = \frac{\Gamma(\sum_b a_b)}{\Gamma(s + \sum_b a_b)} \prod_b \frac{\Gamma(B_i(b) + a_b)}{\Gamma(a_b)} \tag{8}$$

where $s = \sum_b [X(b,i) + Y(b)] = 2n$, i.e. the table sum.

And likewise for $P_i(Q_i \mid H_0)$:

$$P_i(Q_i \mid H_0) = \int_{\bar{p}} P(Q_i, \bar{p} \mid H_0) = \frac{\Gamma(\sum_q a_q)}{\Gamma(s + \sum_q a_q)} \prod_q \frac{\Gamma(Q_i(q) + a_q)}{\Gamma(a_q)} \tag{9}$$

where Q_i (q) is a row sum, i.e. $Q_i(q) = \sum_b X(b,i)$ for q=1 and $Q_i(q) = \sum_b Y(b)$ for q=2.

Then we need to calculate $P_i(B_i, Q_i \mid H_1)$, and this is:

$$P_i(B_i, Q_i \mid H_1) = \int_{\hat{p}} P(B_i, Q_i, \hat{p} \mid H_1) \tag{10}$$

where $\hat{p}$ is a vector of (P(a,1),P(c,1),…P(t,2)).

A conjugate prior for $\hat{p}$ is the Dirichlet distribution:

$$P(\hat{p} \mid a) \sim Dir(a_{a,1}, a_{c,1}, \ldots, a_{t,2}) = \frac{\Gamma(\sum_{b,q} a_{b,q})}{\prod_{b,q} \Gamma(a_{b,q})} \prod_{b,q} P(b,q)^{a_{b,q}-1} \tag{11}$$

where P(b,q)>0 and $\sum_{b,q} P(b,q) = 1$. Given a Dirichlet prior, the joint distribution of B_i, Q_i and $\hat{p}$ is:

$$P(B_i, Q_i, \hat{p} \mid a) = \frac{\Gamma(\sum_{b,q} a_{b,q})}{\prod_{b,q} \Gamma(a_{b,q})} \prod_{b,q} P(b,q)^{T(b,q)+a_{b,q}-1} \tag{12}$$

where $T(b,q) = X(b,i)\delta(q=1) + Y(b)\delta(q=2)$ (δ is Kronecker's symbol).

And the posterior is

$$P(\hat{p} \mid B_i, Q_i, a) \sim Dir(T(b,q) + a_{b,q}) \tag{13}$$

And finally we can calculate:

$$P_i(B_i, Q_i \mid H_1) = \int_{\hat{p}} P(B_i, Q_i, \hat{p} \mid H_1) = \frac{\Gamma(\sum_{b,q} a_{b,q})}{\Gamma(s + \sum_{b,q} a_{b,q})} \prod_{b,q} \frac{\Gamma(T(b,q) + a_{b,q})}{\Gamma(a_{b,q})} \tag{14}$$

If we define $a_b = \sum_q a_{b,q}$, $a_q = \sum_b a_{b,q}$ we are left with:

$$BF_i(H_0; H_1) = \frac{\Gamma(\sum_{b,q} a_{b,q})}{\Gamma(s + \sum_{b,q} a_{b,q})} \prod_b \frac{\Gamma(B(b) + a_b)}{\Gamma(a_b)} * \\ * \prod_q \frac{\Gamma(Q(q) + a_q)}{\Gamma(a_q)} \prod_{b,q} \frac{\Gamma(a_{b,q})}{\Gamma(T(b,q) + a_{b,q})} \tag{15}$$

After calculating $BF_i(H_o; H_1)$ for each i=1,.L (for each column), because we assume independence between columns, we can calculate:

$$BF = \prod_{i=1}^{L} BF_i(H_0, H_1) \tag{16}$$

In addition, for the whole alignment in terms of B and Q (i.e. X and Y) we can calculate the posterior probability of the null model (hypothesis) in the following way:

$$P(H_0 \mid B, Q) = \frac{P(B, Q \mid H_0)P(H_0)}{P(B, Q \mid H_0)P(H_0) + P(B, Q \mid H_1)P(H_1)} \quad (17)$$

Using $P(H_0) = P(H_1) = 0.5$ and the fact that B and Q are independent under the null hypothesis:

$$P(H_0 \mid B, Q) = \frac{\prod_i P_i(B_i \mid H_0) P_i(Q_i \mid H_0)}{\prod_i P_i(B_i \mid H_0) P_i(Q_i \mid H_0) + \prod_i P_i(B_i, Q_i \mid H_1)} \quad (18)$$

Formula (18) can be calculated using (8), (9) and (14).
We can define all priors: $a_{b,q} = 0.5$ as the so-called Jeffreys' prior (the other, more conservative, alternative would be uniform $a_{b,q} = 1$. After comparing the results, we found that Jeffreys' prior gave the best results when we evaluated the method against a series of different alignments (a set of random and biologically derived alignments) using the correlation coefficient as a parameter of success.

The final scores (16) and (18) act as estimates of multiple sequence alignment significance. The alignment of multiple sequences is more significant when BF is small (much smaller then 1) and when $P(H_o \mid B, Q)$ (i.e. $P(H_o \mid X,Y)$) is small (smaller than the probability of the null model for the alignment, i.e. a smaller probability than that the alignment is random). Jeffreys' scale (Jeffreys, 1961) of evidence for Bayes factors is given in Table 2.

Table 2. Jeffrey's scale for the interpretation of Bayes factors (BF) (Jeffreys, 1961).

Bayes Factor (BF) range	Evidence
BF=1	Null hypothesis (model)* is supported
0.3=BF<1	Minimal evidence against null hypothesis (model)*
0.1=BF<0.3	Substantial evidence against null hypothesis (model)*
0.01=BF<0.1	Strong evidence against null hypothesis (model)*
BF<0.01	Decisive evidence against null hypothesis (model)*

* Null model/hypothesis: alignment is random (not biologically relevant).

Background frequencies can be specified by the user or estimated from the input sequence as the observed frequencies of each letter.

Supplementary material 2

for the paper "Quality estimation of multiple sequence alignments by Bayesian hypothesis testing"

A. Tomovic and E. J. Oakeley

List of BF score values, p-values (sFFT, csFFT) for each alignment from the JASPAR database.

Available from: http://www.fmi.ch/groups/functional.genomics/tool-Supp.htm

Supplementary material 3

for the paper "Quality estimation of multiple sequence alignments by Bayesian hypothesis testing"

A. Tomovic and E. J. Oakeley

Table 3. Properties of random and JASPAR alignments

Alignments	Average length	Average number of sequences involved in the alignment
Random alignments	11	32
JASPAR alignments	10	30

The 95% posterior interval for the differences between the mean number of sequences in two groups is [-3.4, 6.8] and the 95% posterior interval for the difference between the mean length of two groups is [-0.6, 1.7]. The posterior intervals were calculated using WinBUGS (Spiegelhalter et al., 2004). Both posterior intervals include 0, therefore there is no significant difference in the average length and the number of sequences within the two groups (random alignments and JASPAR alignments).

WinBUGS code for the calculation of the 95% posterior interval for the difference between the means of the random and JASPAR alignments

```
Model{

for (i in 1:N)
{
  random[i]~dnorm(mu[1],tau[1])
}
```

```
for (i in 1:M)
{
  jaspar[i]~dnorm(mu[2],tau[2])}

for (i in 1:2)
{
 mu[i]~dflat()
 tau[i]~dgamma(1,1)

}
difference<-mu[1]-mu[2]
}
```

Input data:

N- size of vector random (number of random sequences, N=100)

M-size of vector jaspar (number of jaspar alignments, M=107)

random – vector of lengths (or number of sequences) for random alignments

jaspar - vector of lengths (or number of sequences) for jaspar alignments

Supplementary material 4

for the paper "Quality estimation of multiple sequence alignments by Bayesian hypothesis testing"

A. Tomovic and E. J. Oakeley

List of BF score values, p-values (sFFT, csFFT) for each random alignment.

Available from: http://www.fmi.ch/groups/functional.genomics/tool-Supp.htm

Supplementary material 5

for the paper "Quality estimation of multiple sequence alignments by Bayesian hypothesis testing"

A. Tomovic and E. J. Oakeley

Table 1. Methods for the statistical estimation of short ungapped alignments and their running time (L - length of alignment, M - size of the lattice).

Method	**Running time**
NC–naïve method (Hertz and Stormo, 1999)	$O(L^2M^2)$
LD method (Hertz and Stormo, 1999)	O(LMlog(LM))
sFFT (Keich, 2005)	O(LMlog(LM))
csFFT (Nagarajan, et al., 2005)	$O(\sqrt{LM}\log(\sqrt{LM}))$

References

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

Keich, U. (2005) **sFFT: a faster accurate computation of the p-value of the entropy score**, *J Comput Biol, 12, 416-430.*

Nagarajan, N., Jones, N. and Keich, U. (2005) **Computing the P-value of the information content from an alignment of multiple sequences**, *Bioinformatics, 21 Suppl 1, i311-318.*

Supplementary material 6

for the paper "Quality estimation of multiple sequence alignments by Bayesian hypothesis testing"

A. Tomovic and E. J. Oakeley

A web-based implementation of the proposed method is publically available from: http://www.fmi.ch/groups/functional.genomics/tool.htm.

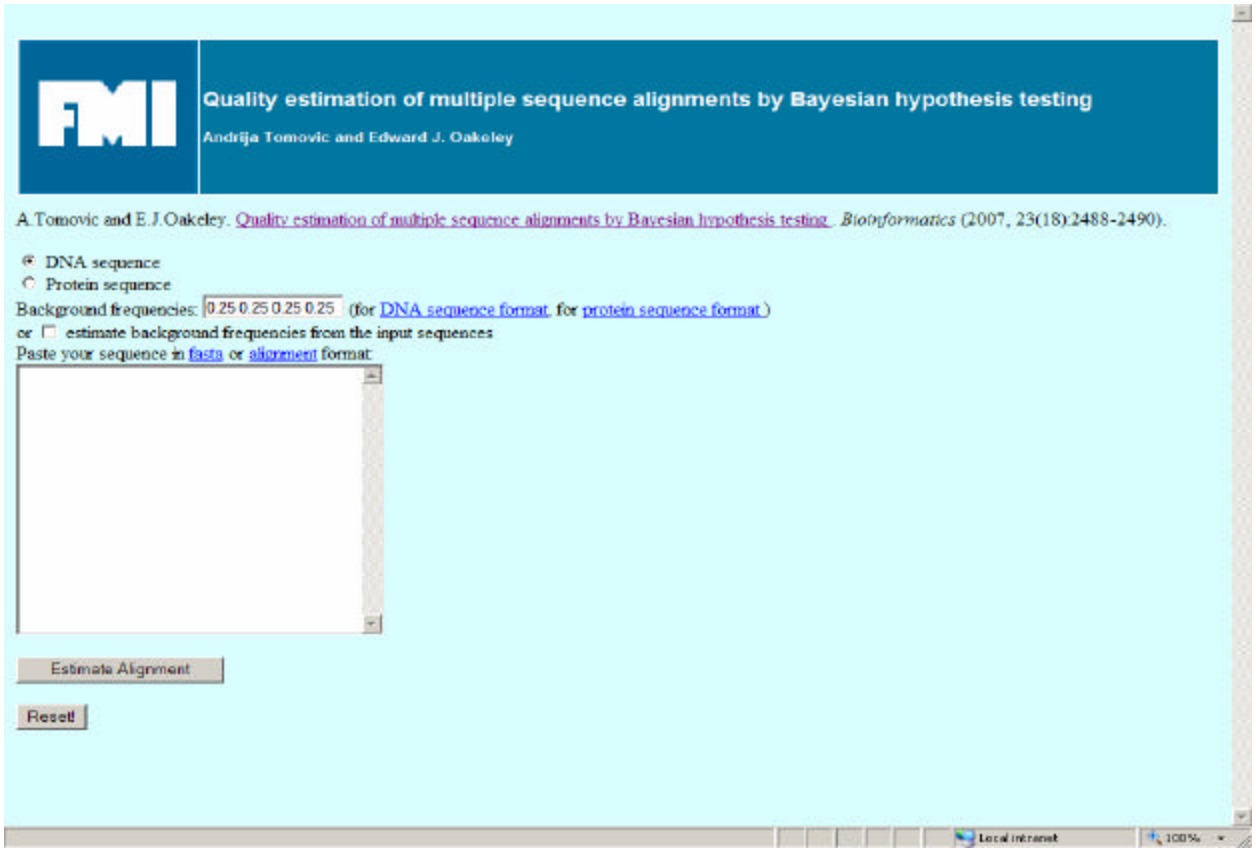


Figure 1. *Web-based tool for the quality estimation of multiple sequence alignments by Bayesian hypothesis testing.*

Acknowledgments

First of all I would like to thank Dr. Edward J. Oakeley for giving me an opportunity to do my thesis in his group at the Friedrich Miescher Institute for Biomedical Research. I appreciate very much his flexibility, support and friendly supervision.

Special thanks to my PhD committee members Prof. Dr. Andreas Engel, Prof. Dr. Torsten Schwede and Prof. Dr. Patrick Mathias for supporting my thesis at the University of Basel. I appreciate very much their advice and monitoring of the progress of my PhD work. Discussions during the PhD committee meetings were very fruitful for my work. In addition, I would like to thank Prof. Dr. Torsten Schwede and all members of his group for allowing me to attend their group meetings and annual retries.

I appreciate very much useful scientific discussion with Michael Stadler and Michael Rebhan. I would like to thank Prof. Dr. Gill Bejerano, Prof. Dr. Frank Hampel, Dr Hans-Rudolf Roth, Prof. Dr. Akinori Sarai, Prof. Dr. Olga Mayans and Dr. Eugene Krissinel for useful discussions and advice.

I want to express my gratitude to my parents, brother and friends (Aleka, Boba, Mirko, Tanja, Gagi, Dragana, …) for their support.

Finally, I want to thank the Novartis Research Foundation for the financial support of the work on this thesis.

Andrija Tomovic

Hallwylstrasse 56, 8004 Zurich, Switzerland
Tel: +41 78 857 4775
e-mail: tomovic.andrija@gmail.com
URL: http://www.fmi.ch/members/andrija.tomovic

Education

2008 • ***PhD in Bioinformatics (Summa Cum Laude; 6.0 out of 6.0); University of Basel, Friedrich Miescher Institute for Biomedical Research, Novartis Research Foundation, Basel, Switzerland;***
PhD thesis: "Computational analysis of promoters and DNA-protein interactions"

2005 • ***MSc in Computer Science, Faculty of Mathematics, University of Belgrade / GPA (grade point average) is 10.00 (out of 10.00)***
Master thesis: "Algorithms for using n-grams in data mining"

2003 • ***Diploma (bachelor with honours) in Mathematics, Faculty of Mathematics, University of Belgrade / GPA (grade point average) is 9.86 (out of 10.00); second ranged in generation (~350 students)/***

Certifications

2008 • **SAS Certified Advanced Programmer Credential for SAS 9** (e-course, passed exam, earned the credential and fulfilled the requirements for certification by SAS)

2008 • **Ingenuity Pathways Analysis Certified Analyst** (successfully completed the Ingenuity Pathways Analysis Certification Program Ingenuity System, Inc; passed exam)

2007 • **SAS Certified Base Programmer Credential for SAS 9** (e-course, passed exam, earned the credential and fulfilled the requirements for certification by SAS)

Work experience

1.01.2009 – present • ***Novartis Pharma AG, Basel, Switzerland***
Modeling & Simulation, Modeler.

1.02.2008 – 31.12.2008 • ***3V-Biosciences Inc. Schlieren, Zurich, Switzerland***

Biotech company; American company based in Palo Alto/ California; full permanent position; denomination: Bioinformatician; kind of work: covering quantitative/computational aspects of research and early development (preclinical phases) of drugs against influenza virus, respiratory syncytial virus and rhinovirus.

1.09. 2007 - 1.02. 2008 • ***Novartis Pharma AG, Basel, Switzerland***
Part-time internship within **Department of Modeling & Simulation**. Projects: PK/binding modeling of antibody + drug interactions; dose-regimen optimization in Phase II

1.03.2004 – 1.01.2008 • ***Faculty of Economics, University of Belgrade***
Department of Mathematics and Statistics
Teaching assistant -
-conducted tutorials on *Mathematics*; *Data Bases* (IBM DB2, SQL, SQL embedded in JAVA); *Programming Language* (JAVA)
-grade report and exams

1.03.2004 – 31.10.2004 • ***Faculty of Mathematics, University of Belgrade***
Visiting teaching assistant –
-conducted tutorials on *Introduction to Programming* (HTML, programming language C and basic algorithms); *Introduction to Computer Systems* (programming language C++ and OOP concepts)
-grade report and exams

1.03.2004 – 31.12.2004 • ***Faculty of Economics, Brcko, Bosnia and Herzegovina University of Sarajevo***
Visiting teaching assistant -
-conducted tutorials on *Data Bases* (Microsoft Access, IBM DB2, SQL)
-grade report and exams

5.08.2003 –31.10.2003 • ***Novartis Pharma AG, Basel, Switzerland***
Cellular Imaging group, Kind of work: design, develop, test, document and validate automated image analysis applications, development of tools and modules as parts of integrated software package - "analySIS / Softver Imaging System", statistical analysis.

12.08.2002 – 28.02.2003 • ***Novartis Pharma AG, Basel, Switzerland***
Working as a trainee: design, develop, test, document and validate automated image analysis applications, development of tools and modules as parts of integrated software package - "analySIS / Softver Imaging System", statistical analysis.

10.06.2000 – 20.07.2002 • ***Television Yugoslavia, YU info channel***,

Belgrade, Yugoslavia (Serbia and Montenegro)
Worked in the computer center of television, managing IT infrastructure and systems

Publications

• Andrija Tomovic, Michael Stadler, Edward J. Oakeley (2009). **Transcription factor site dependencies in human, mouse and rat genomes.** *BMC Bioinformatics 2009, 10:339.*

• M. Stankovic, A. Nikolic, A. Divac, Lj. Rakicevic, A. Tomovic, M. Mitic-Milikic, Lj. Nagorni-Obradovic, M. Grujic, N. Petrovic-Stanojevic, M. Andjelic-Jelic, V. Dopudja-Pantic, D. Radojkovic (2009). **Matrix metalloproteinases gene variants in idiopathic disseminated bronchiectasis.** *J. Investig Med. 2009 Mar;57(3):500-3.*

• M. Stankovic, A. Nestorovic, A. Tomovic, N. Petrovic-Stanojevic, M. Andjelic-Jelic, V. Dopudja-Pantic, Lj. Nagorni-Obradovic , M. Mitic-Milikic, D. Radojkovic (2009). **TNF-a-308 promotor polymorphism in patients with chronic obstructive pulmonary disease and lung cancer.** *Neoplasma Vol.56, No.4, p.348-352, 2009.*

• Andrija Tomovic and Edward J. Oakeley (2008). **Computational structural analysis: Multiple proteins bound to DNA**. *PLoS One,* 2(3):357-62.

• M. Stankovic, A. Nikolic, A. Divac, A. Tomovic, N. Petrovic-Stanojevic, M. Andjelic, V. Dopudja-Pantic, M. Surlan, I. Vujicic, D. Ponomarev, M. Mitic-Milikic, J. Kusic, D. Radojkovic (2008). ***The CFTR M470V Gene Variant as a potential Modifier of Copd Severity.*** *Genetic Testing, 12(3):357-62.*

• Andrija Tomovic and Edward J. Oakeley (2007). ***Quality Estimation of Multiple Sequence Alignments by Bayesian Hypothesis Testing***. *Bioinformatics, 23(18):2488-2490.*

• Andrija Tomovic and Predrag Janicic (2007). **A Variant of N-Gram Based Language Classification.** *Lecture Notes in Artificial Intelligence (sublibrary of Lecture Notes in Computer Science), 4733/2007: 410-421.*

• Andrija Tomovic and Edward J. Oakeley (2007). ***Position Dependencies in Transcription Factor Binding Sites***. *Bioinformatics*, 23(8): 933-41.

• Vesna Cojbasic and Andrija Tomovic (2007). ***Nonparametric Confidence Intervals for Population Variance of One Sample and the Difference of Variances of Two Samples.*** *Computational Statistics and Data Analysis*, 51(12): 5562-5578.

• Andrija Tomovic, Predrag Janicic, and Vlado Keselj (2006). ***n-gram-based Classification and Unsupervised Hierarchical Clustering of Genome Sequences***. *Computer Methods and Programs in Biomedicine*, 81(2): 137-53.

• A. Hagedorn, P-G. Germann, U. Junker-Walker, A. Tomovic, W. Seewald, A. Polkinghorne, A. Posposchil (2005). ***Immunohistochemical study about the Flt-1/VEGFR1 expression in the gastrointestinal tract of mouse, rat, dog, swine and monkey***. *Experimental and Toxicologic Pathology*, 57(2): 149-59.

• Gordana Pavlovic-Lazetic, Nenad Mitic, Andrija Tomovic, Mirjana Pavlovic, Milos Beljanski (2005). ***SARS CoV genome polymorphism: a bioinformatics study.*** *Genomics, Proteomics & Bioinformatics*, 3(1):18-35.

Awards, special grants and duties

• *Ad hoc* reviewer for Bioinformatics, Molecular Diversity, PLOS One

• Selected to participate at the Novartis Biotechnology Leadership Camp (BioCamp), on August 27-29, 2007 (BioCamp provides an opportunity for selected (top 40) university students from all over Europe to learn about the multifaceted world of biotechnology)

• Program Committee Member for 3rd ISCB Student Council Symposium SCS3, Austria, 2007

• Member of ISCB (The International Society for Computational Biology); Member of AAAS (The American Association for the Advancement of Science)

• One of the organizers for the seminar "Basel Bioinformatics Seminar Series", University of Basel, Biozentrum, winter semester 2006/07

• July, 2006 Grant Award (3500 EUR), FREN, Belgrade - Research Project Proposal – "Web-based tool for determination of nonparametric confidence intervals for the one- and two-sample problems", coauthor with Dr. Vesna Cojbasic; second place in competition for the research project proposal grants

• Reviewer/program committee member for the competition "The best technology innovation-2006, -2007, -2008" organized by Government of Republic of Serbia, Ministry of Science

• IAESTE internship Novartis, Basel, Switzerland, 12.08.2002-28.02.2003

• October, 2001 scholarship of the Kingdom of Norway /as one of the best student in Yugoslavia /

• December, 2001 award from Faculty of Mathematics, University of Belgrade /as one of the best student in generation at the Faculty of Mathematics, University of Belgrade/

• 1999-2003 Scholarship Holder of Serbian Ministry of Education

• 1997 The 3-th place in competitions in chemistry in region and participation in Yugoslav Federal Competitions.