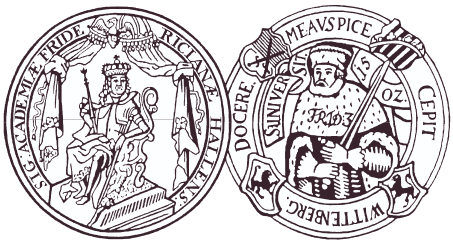
New Approaches for De–novo Motif Discovery Using Phylogenetic Footprinting: From Data Acquisition to Motif Visualization



Dissertation

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr.rer.nat.)

der Naturwissenschaftliche Fakultät III
Agrar- und Ernährungswissenschaften, Geowissenschaften und Informatik
der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

Arthur Martin Nettling

Geb. am 10.06.1982 in Meerane

Gutachter:
1. Prof. Dr. Ivo Grosse
2. Prof. Dr. Peter Stadler

Tag der Verteidigung: 20. April 2017

Acknowledgements

First of all, I thank my beloved wife, Jasmin Nettling, and my whole family for the great support and the patience during the last years. Thank you very much for watching the kids when there was a deadline. Thank you for serving coffee, food, and beer, when I had no time to join lunch or the family party. And thank you for letting me sleep the day when I worked all night again. I am very grateful to my supervisor, adviser, and friend Ivo Grosse, who guided me through my Ph.D. studies. I value very much our honest and passionate discussions at every day or night.

I am also very grateful for the interesting and goal-oriented discussions with Hendrik Treutler. Thank you very much for reading almost everything I have written, for your valuable and honest comments, and your motivating words when evolution behaved again unexpected and unwished.

Further, I thank Andreas Both, Karin Breunig, Jesus Cerquides, Ralf Eggeling, Jan Grau, Jens Keilwagen, Konstantin Kruse, Stefan Posch, Yvonne Pöschel, Marcel Quint, Peter Stadler, and Martin Staege for the valuable discussions, for keeping me on track, and for giving me advice in nearly every Ph.D. related problem. Jan and Jens, you do a really great job with *Jstacs* (`http://www.jstacs.de`). Thank you very much for your kind and fast support every time.

And last but not least, I thank Jörg Weber for helping to develop and implement the web-server `http://difflogo.com` and Charles Bishop for proofreading this thesis.

Peer-reviewed publications

This thesis is a cumulative thesis based on the following publications.

- P Alexiou, T Vergoulis, **M Gleditzsch**, G Prekas, T Dalamagas, M Megraw, I Grosse, T Sellis, AG Hatzigeorgiou. 2009. miRGen 2.0: a database of microRNA genomic information and regulation.
  Nucl. Acids Res. 38 (suppl 1): D137-D141 . *doi:10.1093/nar/gkp888*
- **M Nettling**, N Thieme, A Both, I Grosse. 2014. DRUMS: Disk Repository with Update Management and Select option for high throughput sequencing data.
  BMC bioinformatics, 15:1. *doi:10.1186/1471-2105-15-38*
- **M Nettling**, H Treutler, J Cerquides, I Grosse. 2016. Detecting and correcting the binding-affinity bias in ChIP-Seq data using inter-species information.
  BMC Genomics 17:1. *doi:10.1186/s12864-016-2682-6*
- **M Nettling**, H Treutler, J Cerquides, I Grosse. 2017. Unrealistic phylogenetic trees may improve phylogenetic footprinting.
  Bioinformatics doi: 10.1093/bioinformatics/btx033
- **M Nettling**, H Treutler, J Cerquides, I Grosse. 2017. Combining phylogenetic footprinting with motif models incorporating intra-motif dependencies.
  BMC Bioinformatics, 18:141 *doi: 10.1186/s12859-017-1495-1*
- **M Nettling***, H Treutler*, J Grau, J Keilwagen, S Posch, I Grosse. 2015. DiffLogo: a comparative visualization of sequence motifs.
  BMC bioinformatics, 16:1 *doi:10.1186/s12859-015-0767-x*

I hereby declare that the copyright of the content of the articles Alexiou et al., 2009 and Nettling et al., 2017b is by Oxford University Press. These papers are available at:

- `http://nar.oxfordjournals.org/content/38/suppl_1/D137`
- `https://academic.oup.com/bioinformatics/article/2959846`

I hereby declare that the copyright of the content of the articles Nettling, Thieme, et al., 2014, Nettling, Treutler, Grau, et al., 2015, Nettling et al., 2016, and Nettling et al., 2017a is by the authors. These papers are available at:

- `http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-38`
- `http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0767-x`
- `http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2682-6`
- `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1495-1`

Contents

1 Summary

1.1 English version

The versatility of organisms and their adaptability to environmental changes are essential for their viability and are achieved by expressing proteins on demand. The expression of proteins is orchestrated by the process of gene regulation, which belong to the most complex and comprehensive processes in nature. Hence, the understanding of gene regulation is a prerequisite in modern biology, medicine, and, biodiversity research. A crucial sub-process in gene regulation is the transcriptional initiation, i.e., the interaction of transcription factors (TFs) with their transcription factor binding sites (TFBSs). Hence, predicting TFBSs and their binding motifs in biological sequences is essential for the understanding of gene regulation.

Identifying TFBSs and binding motifs using wet-lab experiments is expensive and time-consuming, and thus neither economical nor feasible. Consequently, bioinformatics approaches have been developed for, first, data acquisition and data preparation, second, for the prediction of putative TFBSs on genomic scale, and third, for the visualization of models for TFBSs.

A typical task in bioinformatics covering these three fields is the prediction of TFBSs in ChIP-sequencing (ChIP-seq) data. This task starts with obtaining sequence data directly from a ChIP-seq experiment or some database. After transforming the data into an appropriate format, *de-novo* motif discovery is performed and putative TFBSs are predicted. Finally, the results are visualized and compared to related work. This thesis covers six peer reviewed articles and one work-in-progress article, which fit into the mentioned three fields as follows.

First, demanded by business and academic needs, the number of data-intensive processes in bioinformatics, like next generation sequencing, is continuously increasing and so is the amount of produced data. The post-processing of these data increases this amount even further. By combining various data sources and different types of data, like sequence data with gene expression data, the data to manage becomes more and more complex. Hence, new databases are needed for an appropriate handling of complex data and new concepts are needed for an efficient handling of increasingly large data volumes.

My colleagues and I developed *miRGen*, a relational *MySQL* database that stores microRNA (miRNA) transcripts with target genes, contained single nucleotide polymor-

phisms (SNPs), TFBSs in near distance, and prominent literature sources. Over the last years, *miRGen* has already become an important resource for researchers that are interested in miRNA regulation and miRNA function.

We also developed the open-source *Java* libarary *DRUMS* which is designed to store billions of position specific DNA related records. *DRUMS* is capable of performing fast and resource sparing requests and runs on a single standard computer. When comparing *DRUMS* to the standard database *MySQL* regarding insert performance and lookup performance on two data sets, it outperforms *MySQL* by a factor of two up to a factor of 15456.

Second, predicting TFBSs in sequence data is essential for the understanding of gene regulation and dozens of bioinformatics approaches have been developed for the prediction of TFBSs. These approaches use diverse statistical characteristics to distinguish TFBSs from their flanking Deoxyribonucleic acid (DNA) and can be subdivided in phylogenetic and non-phylogenetic approaches. Phylogenetic approaches take into account phylogenetic dependencies in aligned sequences of more than one species whereas non-phylogenetic approaches based on sequences of only one species typically take into account intra-motif dependencies. The articles comprising this thesis are related to *de-novo* motif discovery using phylogenetic approaches as follows.

We extended a traditional phylogenetic footprinting model (PFM) by the capability to take into account the binding affinity bias (BA bias) in ChIP-seq data. The BA bias is a result of the over-representation of high-scoring binding sites in ChIP-seq data, causing the inference of potentially distorted motifs. My colleagues and I found that correcting the binding-affinity bias typically leads to softened motifs and significantly improves motif prediction.

We further studied the influence of phylogenetic trees on the performance of *phylogenetic footprinting* and motif prediction. We surprisingly found that unrealistic phylogenetic trees often lead to more accurate predictions of TFBSs than realistic phylogenetic trees.

Based on these results, we developed an approach for *de-novo* motif discovery that extends *phylogenetic footprinting* by the capability of taking into account intra-motif dependencies of higher order. My colleagues and I found intra-motif dependencies of order 1 and 2 in motifs of all investigated species and we found that modelling intra-motif dependencies within *phylogenetic footprinting* significantly improves classification performance.

Third, visualizing the results of motif discovery is fundamental for researchers to interpret, present, and share their findings and sequence logos are the *de facto* standard in biology and bioinformatics to accomplish this task. The number of data sets and motif extraction algorithms is continuously growing and therefor the number of published motifs. Hence, it is often not sufficient to just show motifs but it becomes more and more important

to perceive differences between motifs. Comparing multiple sequence motifs by visual inspection of the corresponding sequence logos can be tricky and especially differences of multiple motifs of the same TF are often hard to perceive.

To address this problem my colleagues and I developed *DiffLogo*, an *R*-package that is specifically designed for visualizing differences between similar sequence motifs. *DiffLogo* visualizes differences between multiple motifs in a tabular plot of all pairwise comparisons. The resulting matrix guides the viewer to the most prominent pairwise differences between motifs. *DiffLogo* is already used in several articles of this thesis to depict differences between motifs of the same TF from phylogenetically related species and to depict differences between motifs of the same TF but captured by different *de-novo* motif discovery approaches.

We know that not all researchers have access to hardware with *R* and *DiffLogo* installed and not all researchers have the time or the technical background to use *R* and *DiffLogo* without high effort. Hence, we integrated *DiffLogo* into the web-server *WebDiffLogo* accessible via `http://difflogo.com`. This web-server allows the user to upload motifs in several common formats. Further, *WebDiffLogo* allows the user to upload motifs of different length and orientation. Hence, *WebDiffLogo* is much easier to use and thus applicable to a much larger community.

Taken together, the findings of this thesis may advance our understanding of transcriptional gene regulation and its evolution. Specifically, our work in the field of data acquisition and data preparation may improve knowledge transfer among researchers and data handling. Our findings in the field of *de–novo* motif discovery based on *phylogenetic footprinting* may lead to an improved prediction of TFBSs. Our work in the field of comparative motif visualisation may help researchers regarding decision making, knowledge sharing, and the presentation of results.

1.2 German version

Die Vielseitigkeit existierender Lebewesen und die Anpassungsfähigkeit an ihre Umwelt ist eine Grundlage für das Leben selbst und ist nur möglich durch die bedarfsbedingte Expression von Proteinen. Die Expression von Proteinen wird durch den Prozess der Genregulation gesteuert, wobei die Genregulation selbst zu den komplexesten und umfangreichsten Prozessen in der Natur zählt. Folglich ist das Verständnis des Genregulationsprozesses sowohl für biologische und medizinische Forschung, als auch für Forschung im Bereich der Biodiversität unabdingbar. Ein entscheidender Teilprozess der Genregulation ist die transkriptionelle Initiation, mit anderen Worten, die Interaktion von Transkriptionsfaktoren (TFen) mit den korrespondierenden Transkriptionsfaktorbindestellen (TFBSen). Die Vorhersage von TFBSen und die Inferenz ihrer Bindemotive ist somit eine unabdingbare Grundlage um den gesamten Prozess der Genregulation zu verstehen.

Die Identifikation von TFBSen und ihren Bindemotiven mittels klassischer Laborexperimente ist jedoch teuer und zeitaufwändig und damit weder ökonomisch noch praktikabel. Folglich wurden bioinformatische Methoden und Ansätze entwickelt, um Daten effizient zu beschaffen und vor zu verarbeiten, um mögliche TFBSen genomweit vorherzusagen und um Modelle für TFBSen zu visualisieren.

Eine typische bioinformatische Aufgabe, welche diese drei Bereiche umfasst, ist die Vorhersage von TFBSen in ChIP-seq Daten. Diese Aufgabe beginnt mit der Akquisition von Sequenzdaten entweder direkt aus einem ChIP-seq Experiment oder aus entsprechenden Datenbanken. Nachdem die Daten aufbereitet und in ein passendes Format transformiert wurden, kann die eigentliche Motivvorhersage beginnen. Abschließend werden die Ergebnisse typischerweise visualisiert und mit denen ähnlicher Arbeiten verglichen. Die vorliegende Dissertation umfasst sechs begutachtete Publikationen und ein Manuskript, welches noch in Arbeit ist, die sich folgendermaßen in die genannten drei Bereiche eingliedern.

Erstens, aufgrund industriellen und akademischen Bedarfs steigt die Zahl der datenintensiven Prozesse in der Bioinformatik an. Ein Beispiel dafür ist Next Generation Sequencing. In gleichem Maße wächst der Umfang produzierter Daten. Die Nachverarbeitung dieser Daten steigert die Datenmenge nochmals. Des Weiteren wird durch die Kombination verschiedener Datenquellen und Datentypen, wie Sequenzdaten und Expressionsdaten, die Komplexität der zu verwaltenden Daten stetig erhöht. Damit steigt der Bedarf an neuen Datenbanken, die in der Lage sind, komplexere Daten zu verwalten. Außerdem werden neue Konzepte benötigt um die immer größer werdenden Datenmengen effizient verwalten zu können.

In diesem Kontext haben meine Kollegen und ich *miRGen* entwickelt, eine relationale *MySQL* Datenbank zur Speicherung von microRNA (miRNA) Transkripten, angereichert mit deren Zielgenen, mit enthaltenen Einzelnukleotid-Polymorphismen (SNPs), mit TFBSen in direkter Umgebung und mit prominenten Literaturquellen. *miRGen* ist bereits zu einer wichtigen Ressource für Forscher geworden, die an der Regulation und der Funktion von miRNAs interessiert sind.

Des Weiteren haben wir die open-source *Java* Bibliothek *DRUMS* zur Speicherung von Milliarden von Datensätzen entwickelt, welche sich positionsspezifisch auf Sequenzen beziehen, wie es bei z. B. SNPs der Fall ist. *DRUMS* ist in der Lage Anfragen schnell und ressourcenschonend zu beantworten und läuft auf Standard-Desktop-Hardware. Bei dem Vergleich von *DRUMS* mit der Standarddatenbanklösung *MySQL* bezüglich Einfügegeschwindigkeit und Anfragegeschwindigkeit ist *DRUMS* 2 bis 15456 mal schneller als *MySQL*.

Zweitens, die Vorhersage von TFBSen in Sequenzdaten ist unabdingbar für das Verständnis des Genregulationsprozesses. Dutzende bioinformatische Ansätze existieren, um dies zu bewerkstelligen. Diese Ansätze nutzen verschiedene statistische Eigenschaften, um TFBSen von flankierender DNA zu unterscheiden. Phylogenetische Ansätze verwenden phylogenetische Abhängigkeiten in alignierten Sequenzen mehrerer Spezies, wohingegen nicht-phylogenetische Ansätze basierend auf Sequenzen von nur einer Spezies normalerweise Nukleotidabhängigkeiten innerhalb des Motivs berücksichtigen können. Die Arbeiten dieser Dissertation sind fokussiert auf die Motiverkennung unter Verwendung phylogenetischer Ansätze.

In diesem Kontext haben wir als Erstes ein traditionelles Phylogenetic Footprinting Modell um die Fähigkeit erweitert den Bindeaffinitätsbias (BA bias) von TFen in ChIP-seq Daten zu berücksichtigen. Der BA bias resultiert aus der Überrepräsentation von hochqualitativen Bindestellen in ChIP-seq Daten und verursacht die Vorhersage von potentiell verzerrten Motiven. Wir konnten zeigen, dass das Korrigieren des BA bias in der Regel zu weicheren Motiven führt und die Motivvorhersage signifikant verbessert.

Des Weiteren haben meine Kollegen und ich den Einfluss phylogenetischer Bäume auf die Leistung von Phylogenetic Footprinting und Motivvorhersage untersucht. Überraschenderweise haben wir entdeckt, dass unrealistische phylogenetische Bäume oftmals zu genaueren Vorhersagen von TFBSen führen als realistische phylogenetische Bäume.

Aufbauend auf dieser Erkenntnis haben wir einen Ansatz zur Motiverkennung entwickelt, welcher Phylogenetic Footprinting um die Fähigkeit erweitert Nukleotidabhängigkeiten höherer Ordnung innerhalb eines Motivs zu modellieren. Wir haben Nukleotidabhängigkeiten erster und zweiter Ordnung in Motiven aller untersuchten Spezies gefunden und wir konnten zeigen, dass das Modellieren von Nukleotidabhängigkeiten im Rahmen von Phylogenetic Footprinting die Vorhersagegüte signifikant verbessert.

Drittens, die Visualisierung der Modelle, die während der Motiverkennung generiert werden ist für Wissenschaftler fundamental um zum einen ihre Ergebnisse selbst interpretieren zu können und zum anderen um Erkenntnisse zu präsentieren und zu teilen. In vielen Bereichen der Biologie und Bioinformatik werden dafür Sequenzlogos verwendet. Durch die stetig steigende Zahl an verfügbaren Datensätzen und Algorithmen zur Motiverkennung wächst die Zahl der veröffentlichten Sequenzmotive. Damit ist es oft nicht mehr ausreichend Sequenzmotive lediglich zu präsentieren bzw. zu visualisieren, sondern es wird immer wichtiger auch Unterschiede zwischen Sequenzmotiven hervorzuheben. Der Vergleich mehrere Sequenzmotive mittels Sequenzlogos kann sich als äußerst schwierig erweisen und im Besonderen ist es auf diese Weise kaum möglich Unterschiede zwischen Motiven des gleichen TFs aus z. B. unterschieldichen Zelllinien zu erkennen.

Meine Kollegen und ich haben dieses Problem mit der Entwicklung von *DiffLogo* adressiert, ein speziell für die Visualisierung von Motivunteschieden entwickeltes R-Paket. *DiffLogo* visualisiert Unterschiede mehrerer Motive in einer tabellarischen Darstellung aller paarweisen Vergleiche. Die resultierende Visualisierung hebt prominente, paarweise Unterschiede farblich hervor und fokussiert somit den Betrachter auf das Wesentliche. *DiffLogo* wird bereits in mehreren Publikationen dieser Dissertation verwendet.

Da nicht alle Wissenschaftler Zugang zu entsprechender Hardware mit installiertem *R* und *DiffLogo* haben und außerdem viele Wissenschaftler nicht genügend Zeit oder technische Erfahrung haben um *R* und *DiffLogo* ohne Probleme zu verwenden, haben wir *DiffLogo* in einen WebServer integriert. Dieser ist über `http://difflogo.com` erreichbar. Der Nutzer kann Sequenzmotive in verschiedenen Formaten hochladen. Außerdem dürfen die Motive unterschiedlich lang sein und eine unterschiedliche Orientierung haben. *DiffLogo* versucht in diesen Fällen die Sequenzmotive zu alignieren. `http://difflogo.com` ist wesentlich einfacher zu verwenden als *DiffLogo* und damit für eine größere Nutzerschaft zugänglich.

Abschließend kann ich sagen, dass die Erkenntnisse dieser Arbeit unser Verständnis der transkriptionellen Genregulation und deren Evolution voranbringen können. Im Detail: Unsere Arbeit und unsere Erkenntnisse im Bereich der Datenakquisition und Datenvorbereitung können den Wissenstransfer zwischen Wissenschaftlern und das allgemeine Datenmanagement verbessern. Unsere Erkenntnisse im Bereich der Motivvorhersage mittels Phylogenetic Footprinting können zu einer verbesserten Vorhersage von TFBS führen. Unsere Arbeit im Bereich der vergleichenden Darstellung von Sequenzmotiven kann Wissenschaftlern bei der Entscheidungsfindung, beim Wissenstranfer, bei der Dokumentation und Präsentation ihrer Ergebnisse helfen.

2 Introduction

From 1857 to 1864, Gregor Mendel studied the inheritance patterns in pea plants and suggested the idea of the existence of discrete inheritable units. At the same time Darwin published his famous work "On the Origin of Species," proposing continual evolution of species (Darwin, 1859). It took over 50 years (1909) until Wilhelm Johannsen coined the word *gene* to name those inheritable units proposed by Mendel, and it took another 44 years before James Watson and Francis Crick published their model for DNA, which is now known as the double-helix model of DNA structure (Watson et al., 1953). Three years later, Francis Crick stated the **central dogma of molecular biology** for the first time. This dogma describes the relationship between DNA, Ribonucleic acid (RNA), and proteins and was finally published in 1970 (F. Crick et al., 1970). Ever since then, molecular biology has undergone a rapid and extensive development and after 150 years, terms like gene, DNA, RNA, proteins, evolution, and mutation have found its way into our daily language.

This development was accompanied by the digital revolution, starting with the invention of the transistor in 1947, the fundamental building block of any modern digital device. In the 70s, home computers were introduced and the transformation of analog to digital data began. In 1969 the first message was sent over the ARPANET, the predecessor of the Internet, which became publicly accessible in 1991 as the World Wide Web. Nowadays, 50% of the world population has access to the Internet[1] and everybody is passively and actively generating data that can be used to improve our daily lives. For example, GPS data of many individuals can be used to predict traffic anomalies (Pang et al., 2013), data of fitness trackers can be used to prevent cardiovascular diseases (Neubeck et al., 2015), and differences in human genomes can be used to understand the genetic contribution to various diseases (E. P. Consortium et al., 2012; 1. G. P. Consortium et al., 2012; Sudmant et al., 2015). It is expected that in 2020 the digital universe will comprise over $5,200$ gigabytes per living person, summing up to 40 trillion gigabytes of data (Gantz et al., 2012; Dragland, 2013). The continuously growing quantity of information and its availability at any time, in any place, already permeates both our work and our daily lives.

The digital revolution also enabled new data sources and new technologies in the field of molecular biology, e.g., sequence data and sequencing technologies. Starting in 1990, it took 13 years to sequence the first human genome, whereas in 2015, this task could be accomplished in 26 hours (Miller et al., 2015). The extensive and continuously growing amount of data enabled the emergence of new sciences like bioinformatics, unleashing

[1] `http://www.internetworldstats.com/stats.htm`

new potentials regarding the study of fundamental biological processes like the complex process of gene regulation. Nowadays, molecular biology is a field of research that serves a broad audience with a scientific background as well as non-scientific backgrounds. In this sense, with this thesis, I want to contribute to a deeper understanding of the process of gene regulation and its evolution. Specifically, my colleagues and I try to contribute to data acquisition and data preparation by developing new databases for knowledge sharing and for efficient data handling. We also attempt to develop new approaches based on *phylogenetic footprinting* for the *de–novo* motif discovery in ChIP-seq data. Finally, we try to develop a new approach for the comparative visualization of sequence motifs. The next four sections introduce the reader to molecular biology, computer science, bioinformatics, and the research objectives of this thesis.

2.1 Biological background

This section gives a general introduction into gene expression and gene regulation. The introduction also includes a description of the transcriptional initiation and the post–transcriptional gene regulation by miRNAs.

2.1.1 Gene expression and gene regulation

DNA is the intra–cellular substance that carries the definition of an organism's potentials. It is composed of the nucleotides of the four organic bases adenine (A), cytosine (C), guanine (G), and thymine (T). The sequence of these bases encodes the genetic information. Genes are information units on the DNA and gene expression is the process used by all known life that translates genes into proteins or functional RNA. In all organisms, gene expression comprises at least two processes, transcription and translation (Figure 2.1). Transcription is the process that transcribes a DNA sequence to the corresponding messenger RNA (mRNA) sequence. Translation is the process that translates the mRNA into the corresponding polypeptide, which may then fold to a protein.

Gene regulation is the process that regulates gene expression and is the basis for cellular differentiation, morphogenesis, versatility, and adaptability of any organism. Whenever a protein is needed, due to e.g., environmental stimuli, a complex signaling cascade is initiated to first, make the corresponding DNA region accessible for the transcription machinery and second, to ensure the correct processing of the genetic information. This fundamental process is established by diverse sub–processes such as epigenetics (Reik, 2007; Slotkin et al., 2007; Dolinoy et al., 2007), regulation by miRNAs (He et al., 2004; K. Chen et al., 2007), siRNA interference (Fougerolles et al., 2007; Tam et al., 2008), and alternative splicing (Sultan et al., 2008; Luco et al., 2010).

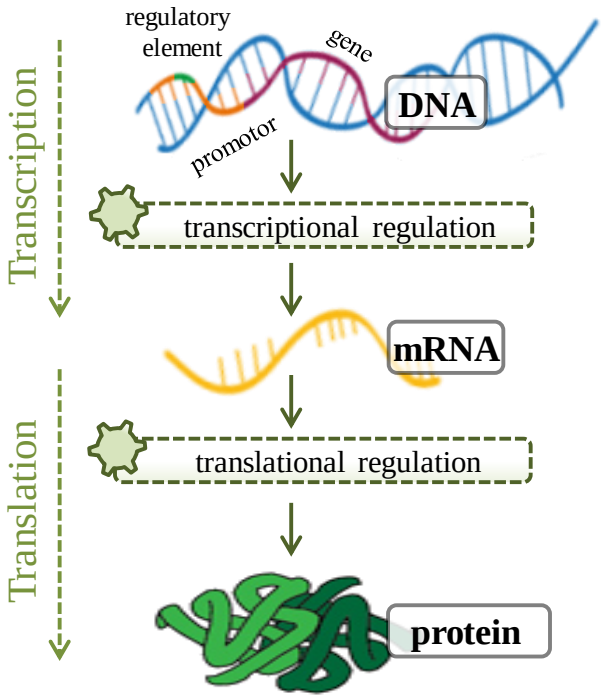


Figure 2.1: Flowchart of gene expression. A gene is an information unit on the DNA. The process of gene expression typically comprises the two sub–processes transcription and translation. Transcription is the process that transcribes the gene to the corresponding mRNA. Translation is the process that translates the mRNA into the corresponding protein.

2.1.2 Transcriptional initiation

An important sub–process is the initiation of transcription, which is mediated by the interaction of TFs with the DNA. Specifically, TFs are proteins that bind to specific DNA signals, so called TFBSs. TFBSs are often located in the upstream promotor region of a gene but can also be found in other, more distant, intergenic regions. The binding of a TF to a TFBS can enhance or represses the expression of the gene. **Figure 2.2** shows a schematic representation of the binding of three TFs to the DNA to initiate the transcription of the downstream located gene.

Uncovering TFBSs in genomic DNA and inferring DNA binding motifs for TFs, also known as *de–novo* motif discovery, is a prerequisite in modern biology, medicine, and biodiversity research (D'haeseleer, 2006). TFBSs cover short DNA regions of about 10 bases in contrast to the human genome which is about 3.3 billion bases long (Tompa et al., 2005). The complexity of *de–novo* motif discovery even doubles due to the double strand property of the DNA that allows TFBSs to be located on either of the two reverse complementary DNA strands. Hence, studying gene regulation by time-consuming and expensive wet-lab exper-

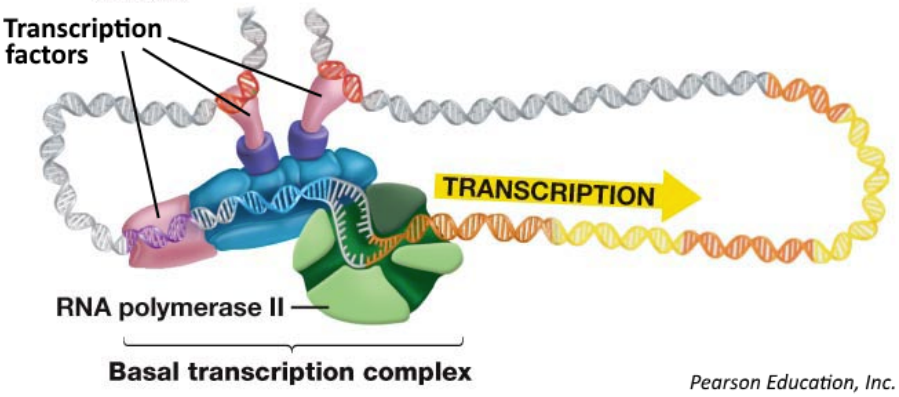


Figure 2.2: Initiation of transcription. Three transcription factors bind to the DNA to mediate the binding of the basal transcription complex and hence the start of transcription. The gene to transcribe consists of three exons (red) and two introns (yellow).

iments is neither economical nor practical, and the computational investigation of DNA binding motifs and their binding sites seems to be feasible. New approaches for uncovering TFBSs in genomic DNA using *phylogenetic footprinting* are studied in **Sections 3.2.2** and **3.2.3** and in **Sections 5.2** and **5.3**.

2.1.3 Gene regulation by miRNAs

miRNAs are short single stranded non-coding RNAs with a length of about 22 bases. These post–transcriptionally influence gene expression by binding to specific sites within the 3'-untranslated region (UTR) of mRNAs, causing a decrease of gene expression by inhibiting the translation of mRNAs or by directly causing degradation of mRNAs. miRNAs appear to target about 60% of the human genes and other mammals and hence play a key role in the development of organisms (Carrington et al., 2003).

Especially in medicine, the understanding of gene regulation by miRNAs is of great interest since these have been linked to several human pathologies such as cardiovascular and neurodegenerative diseases as well as human malignancies (Calin et al., 2006; Nelson et al., 2008). Further, miRNAs are believed to be involved in many stages of cancer progression by both promoting and suppressing oncogenesis, tumor growth, invasion, and metastasis (Farazi et al., 2011; Small et al., 2011). The influence of miRNAs on gene regulation is studied in **Sections 3.1.1** and **4.1**.

2.2 Computer science background

This section gives a general introduction to the programming techniques and concepts used in this thesis. Specifically, the first subsection will introduce the reader into Java and the open source *Java* library *Jstacs*. The second subsection gives a short overview about *R*. In the third subsection, the reader finds a brief overview to relational and non relational databases.

2.2.1 The *Java* programming language and the *Java* library *Jstacs*

Java is one of the most popular programming languages in use (O'Grady, 2015). It is concurrent, class-based, and object-oriented (Gosling et al., 2014). *Java* code needs to be compiled into standard byte code before it can be executed on all platforms that support *Java*. Regarding this thesis, we use *Java* to implement *DRUMS* (**Sections 3.1.2** and **4.2**) and new approaches for *de–novo* motif discovery based on *phylogenetic footprinting* (**Sections 3.2.1-3.2.3** and **5.1-5.3**).

Jstacs is an open source *Java* library for the statistical analysis of biological sequences developed by the groups *Pattern Recognition and Bioinformatics at the Institute of Computer Science of Martin Luther University Halle-Wittenberg* and the *Bioinformatics group of the Julius Kuehn Institute* (Grau, Keilwagen, et al., 2012). *Jstacs* provides convenient and efficient classes for the representation of sequence data, many statistical models suitable for the prediction of TFBSs in sequence data, ready to use numerical optimization procedures, and several performance measures. The design of *Jstacs* is a strictly object-oriented framework with a deep class hierarchy and there is a rich documentation. Hence, *Jstacs* is easy-to-use, extensible, and customizable to a great extend.

In this thesis, my colleagues and I use *Jstacs* to implement prototypes of new approaches for *de–novo* motif discovery based on *phylogenetic footprinting* (**Sections 3.2.1-3.2.3** and **Sections 5.1-5.3**). Specifically, we extend the class `Sample` from *Jstacs* by the capability to handle multi-dimensional sequence data (e.g. alignments). We the resulting class `PhyloSample` also provides convenient methods for the processing of phylogenetic data. Further, we implement several `Models` related to *phylogenetic footprinting* extending *Jstacs*' `AbstractModel` class. This allows us to use optimization procedures and performance measures available in *Jstacs* and thus facilitates development and decreases the probability of implementation bugs. Finally, we extend *Jstacs*' numerical optimization procedures by the capability to optimize parameters from phylogenetic trees and PFMs. We make our implementations available in the *PhyFoo* project on GitHub[1]. **Figure 2.3**

[1] `https://github.com/mgledi/PhyFoo`

shows a high-level overview regarding the usage of *Jstacs* for prototyping new approaches for *de–novo* motif discovery based on *phylogenetic footprinting*.

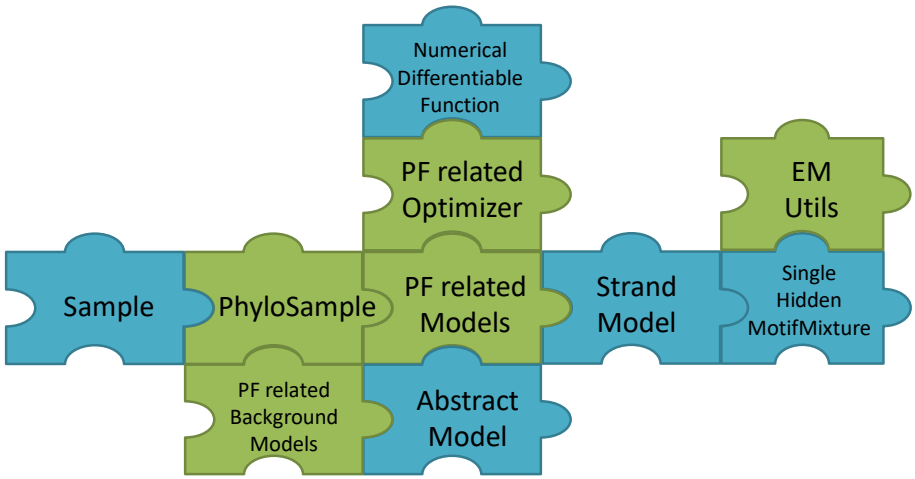


Figure 2.3: High level overview regarding the usage of *Jstacs* within the *PhyFoo* project. Blue puzzle pieces denote *Jstacs* classes. Green puzzle pieces denote parts of the *PhyFoo* project.
The `phylogenetic footprinting (PF) related models` in the *PhyFoo* project extend *Jstacs*' `AbstractModel` to allow the usage of *Jstacs*' optimization procedures and performance measures. This also includes the possibility to wrap `PF related models` in *Jstacs*' `StrandModel` to allow modelling TFBSs on both DNAs strands. Finally, *Jstacs*' `SingleHiddenMotifMixture` class comprises methods to run an EM algorithm on the `PF related models`. All `PF related models` need a `PhyloSample` as input whereas the `PhyloSample` class extends *Jstacs*' `Sample` class by the capability to handle multi-dimensional sequence data.

2.2.2 The *R* programming language and Bioconductor

R is a programming language and software environment for statistical computing. In contrast to Java, *R* is an interpreted language, i.e., *R* code does not need to be compiled and hence can be executed at any time, allowing the developer to easily and quickly prototype new computational methods. Another reason for the continuously growing popularity of *R* among researchers is that many libraries exist providing, inter alia, a wide variety of statistical and graphical techniques (Tippmann et al., 2015). At the end of 2016 the Comprehensive R Archive Network (CRAN) package repository features 9699 available packages. Another interesting feature for advanced users is that they can manipulate *R* objects directly using other, more efficient programming languages like *Java*, *C*, *C++*, or *Phyton*.

R is extensively used in the field of bioinformatics and *Bioconductor* is an open development software project that provides many high quality *R* packages regarding this field (R. C. Gentleman et al., 2004). Specifically, the *Bioconductor* project provides over 1000 powerful statistical and graphical packages for the analysis of genomic data. Popular examples are the package *genefilter* for the filtering of genes from high–throughput data (Bourgon et al., 2010), the package *GenomicAlignments* for the handling of short genomic alignments (M. Lawrence et al., 2013), and the package *seqLogo* for the visualization of sequence motifs (Bembom, n.d.). It also provides over 900 packages with annotation data for, e.g., human, mouse, yeast, or rockcress (Huber et al., 2015). A popular example in this group is the *biomaRt* package which integrates BioMart data resources (e.g. Ensembl) with data analysis software in *Bioconductor* (Durinck et al., 2005). Further, the *Bioconductor* project contains more than 300 packages providing extensive experimental data of any kind, e.g., sequencing data or expression data.

I also want to mention the two projects *Shiny*[1] and *OpenCPU*[2] that allow the integration of *R* into a web-server and hence make it easy to publish and share work with other researchers.

Regarding this thesis, *R* is used to implement *DiffLogo* (**Sections 3.3.1, 3.3.2** and **6.1**). *DiffLogo* is available as *R* package in the *Bioconductor* software suite[3] and via GitHub[4].

2.2.3 Databases

Databases are used to store collections of data. The data are typically organized in a structured way enabling fast and purposeful access. There exist hundreds of different databases that can be divided into two groups.

The first group, relational databases, comprises databases that are based on the relational model of data (Codd, 1970). That means, that data is organized in tables consisting of rows and columns, where each row can be identified with a unique key. Data in a relational database can be queried using structured query language (SQL). Relational databases are used when the structure of the data is well defined, i.e., rely on a schema, and when the data is mainly accessed using complex queries with many relations. Examples for relational databases are *MySQL*, *DB2*, or *PostgreSQL*. Regarding this thesis, *MySQL* is used for the implementation of *miRGen* (**Sections 3.1.1** and **4.1**).

The second group, not only SQL (NoSQL) databases or non relational databases, are often simpler designed than relational databases and typically lack tabular relations. The

[1] `https://shiny.rstudio.com/`
[2] `https://www.opencpu.org/`
[3] `http://bioconductor.org/packages/release/bioc/html/DiffLogo.html`
[4] `https://github.com/mgledi/DiffLogo`

data stored in NoSQL databases can be unstructured, i.e., NoSQL databases are typically schema-free. Hence, these databases are typically faster, can store more data, have a higher scalability, and are easier to maintain. Examples for NoSQL databases are *MongoDB*, *Cassandra*, or *BerkleyDB*.

Sometimes, it is not sufficient to just choose a meaningful database or database management system for a certain data handling problem. In these cases a comprehensive storage concept is needed. One example for such a storage concept is the Disk Repository with Update Management (*DRUM*) concept which was initially proposed by Lee et al. to store billions of URLs with meta-data using a single-server implementation (Lee et al., 2009). The central idea of the *DRUM* concept is to maintain fast sequential read and write access from and to the underlying storage device by holding and preparing as much records as possible in memory.

In context of this thesis, my colleagues and I propose a NoSQL database based on the *DRUM* concept for the management of large biological datasets on single desktop hardware (**Sections 3.1.2** and **4.2**).

2.3 Bioinformatics background

This section introduces the reader into the fields of bioinformatics touched by this thesis and its limitations. Specifically, this section includes a brief description of integration of biological data (**Section 2.3.1**) and ChIP-seq data analysis (**Section 2.3.2**). The reader will also be introduced to the idea of *de–novo* motif discovery based on *phylogenetic footprinting* (**Section 2.3.3**) and to the visualization of sequence motifs (**Section 2.3.4**).

2.3.1 Integration of biological data

With the continuously growing amount of biological data, the need to store, share, and organize it also grows. The current NAR database issue comprises 62 articles describing new biological databases and 112 articles describing updates on existing databases for e.g., storing ChIP-seq data (Daniel J Rigden et al., 2016). The online molecular biology database collection therewith now comprises 1685 biological databases (Daniel J. Rigden et al., 2016).

Biological databases can be divided into primary and secondary databases. Primary databases typically contain data of only one type. Their main purpose is completeness and up-to-dateness. Secondary databases often combine data from primary databases and typically already analyze the data depending on the corresponding requirements.

My colleagues and I identified two limitations regarding biological databases. First, there exists no database that provides comprehensive information about miRNA transcripts together with their regulation by transcription factors, expression profiles, SNPs, and miRNA targets. We address this problem in **Sections 3.1.1** and **4.1**. Second, there exists no database that is capable of storing billions of position-specific DNA-related records, performing fast and resource saving requests, and running on a standard personal computer. We propose a database which fulfills these requirements and we present our idea in **Sections 3.1.2** and **4.2**.

2.3.2 ChIP-seq data analysis

ChIP-seq is a powerful experimental method for identifying binding sites for TFs and other proteins on a genomic scale (T. Bailey et al., 2013). The idea is to immunoprecipitate the DNA-bound protein using a specific antibody. The bound DNA is then coprecipitated, purified, and sequenced resulting in extremely large sets of raw data which necessitate different post–processing steps like quality control, read mapping, and peak detection. **Figure 2.4** gives a high level overview over a ChIP-seq experiment.

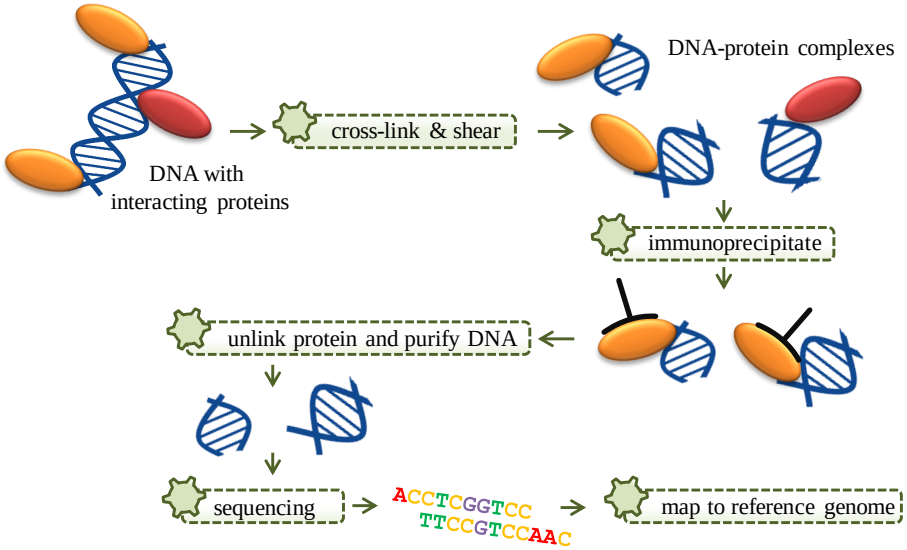


Figure 2.4: High level overview of a ChIP-seq experiment. The ChIP-seq experiment starts with crosslinking a protein to DNA. The DNA-protein complexes are sheared by, e.g. sonication, and immunoprecipitated using a specific antibody. Next, the proteins are unlinked and the DNA fragments are purified. The purified DNA is sequenced and the resulting millions of short DNA sequences are mapped against a reference genome.

The filtered ChIP-seq data is then typically subjected to *de–novo* motif discovery and dozens of approaches exist for this purpose (Tran et al., 2014). Unfortunately, as many

other techniques, motifs predicted by these computational approaches are distorted by the presence of various biases, such as the ubiquitous binding-affinity bias (Håndstad et al., 2011; Ross et al., 2013; Timothy L. Bailey, Krajewski, et al., 2013). My colleagues and I propose an approach to estimate and diminish the BA bias in ChIP-seq data. We present our idea in **Sections 3.2.1** and **5.1**.

2.3.3 *De–novo* motif discovery based on *phylogenetic footprinting*

The last decade has witnessed a spectacular development of sequencing technologies unleashing new potentials in identifying TFBSs (D. S. Johnson et al., 2007; I. V. Kulakovskiy et al., 2010; Furey, 2012). Countless approaches exist for predicting TFBSs of known TFs and for *de–novo* motif discovery of TFBSs in sequence data. There are many meaningful ways to group these approaches. Tran et al. (Tran et al., 2014) grouped several motif finding web tools by the way the sequence motif is modelled. The resulting groups are *Profiles*, consensus sequences (*Consensuses*), *Projections*, *Graph representations*, *Clustering* of k-mers, and *Tree-based* data structures. Another way to distinguish different approaches could be by learning principle like *generative learning principles* and *discriminative learning principles*. A list with 36 different tools for motif discovery is available in the supplement of the work of Zambelli et al. (2012). In this thesis, we divide approaches for *de–novo* motif discovery into phylogenetic approaches and non-phylogenetic or single-species approaches.

Due to the increasing number of available genomes from different organisms and due to ever-increasing computational resources, approaches that incorporate sequence information from phylogenetically related species have become increasingly attractive. These approaches can typically be assigned to phylogenetic footprinting or phylogenetic shadowing. The border between both is very blurry as phylogenetic footprinting is called phylogenetic shadowing when a large number of closely related species is used.

The idea behind phylogenetic footprinting and phylogenetic shadowing is that regions containing functional elements, like TFBSs, are considered to evolve more slowly than regions without any functional elements. The reason for this is that mutations in functional elements are more likely to be lethal than mutations in non-functional sequences. Thus, we observe manifested mutations more often in regions without any functional elements than in regions comprising functional elements. With other words, sequences comprising functional elements are subject to a larger evolutionary pressure than sequences without functional elements. To profit from this idea, approaches incorporating sequence information from phylogenetically related species typically use alignments of orthologous sequences as well as a phylogenetic tree that represents the relationship among the species of interest. **Figure 2.5** illustrates this idea.

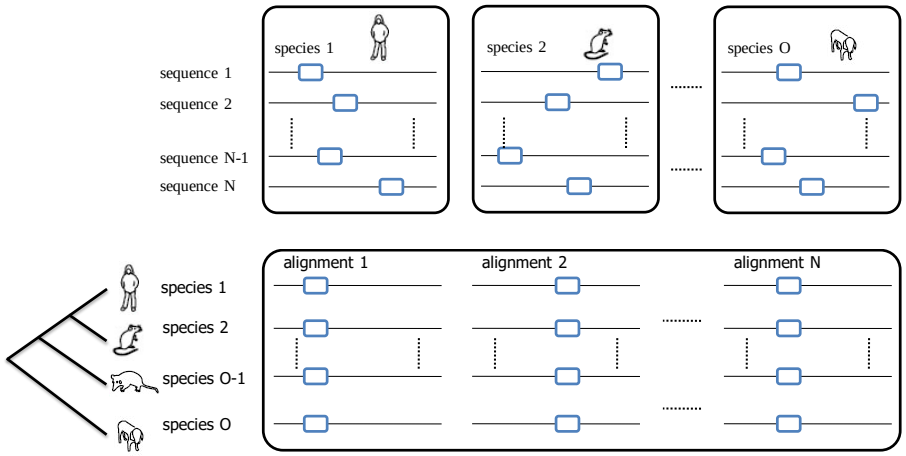


Figure 2.5: Idea of phylogenetic footprinting. Sequences are denoted by black lines and binding sites are depicted by blue boxes. The **upper panel** illustrates the idea of *de–novo* motif discovery ignoring dependencies among orthologous sequences from phylogenetically related species. Orthologous sequences are arranged in rows. All sequences among all species are assumed to be statistically independent from each other.
The **lower panel** illustrates the idea of *de–novo* motif discovery incorporating dependencies among orthologous sequences from phylogenetically related species. Each set of orthologous sequences is now aligned. The sequences of an alignment are phylogenetically related, i.e., statistically dependent. The phylogenetic context is given by the phylogenetic tree on the left.

The above described idea helps methods incorporating phylogenetic dependencies to detect functional elements like TFBSs with higher sensitivity compared to methods neglecting phylogenetic dependencies (Moses et al., 2004; Gertz et al., 2006). Several tools using alignments of orthologous sequences have been proposed to uncover TFBSs, e.g., *FootPrinter* (Blanchette et al., 2003), *PhyME* (Sinha et al., 2004), *MONKEY* (Moses et al., 2004), *PhyloGibbs* (Siddharthan et al., 2005), *Phylogenetic Gibbs Sampler* (Newberg et al., 2007), *PhyloGibbs-MP* (Siddharthan, 2008), and *MotEvo* (Arnold et al., 2012).

All of these approaches use a phylogenetic tree with predefined substitution probabilities as preliminary input, but none of them investigates the influence of different phylogenetic trees and different substitution probabilities on classification performance and motif prediction on ChIP-seq data. My colleagues and I fill this knowledge gap in **Sections 3.2.2** and **5.2**. Another limitation of these approaches is that they neglect intra–motif dependencies, although it has been shown that more complex motif models that take into account intra–motif dependencies outperform simpler motif models like the position weight matrix (PWM) model. We address this problem in **Sections 3.2.3** and **5.3**.

2.3.4 visualization of sequence motifs

An important task in research is the visualization of results and sequence logos are the *de facto* standard for visualizing sequence motifs obtained from *de–novo* motif discovery (Schneider et al., 1990). A sequence logo represents the characteristics of each motif position by the two properties stack height and symbol height within a stack. The stack height is proportional to the information content of the symbol distribution and the symbol height is proportional to the degree of symbol abundance. **Figure 2.6** shows an example of two similar sequence logos of the TF CTCF.

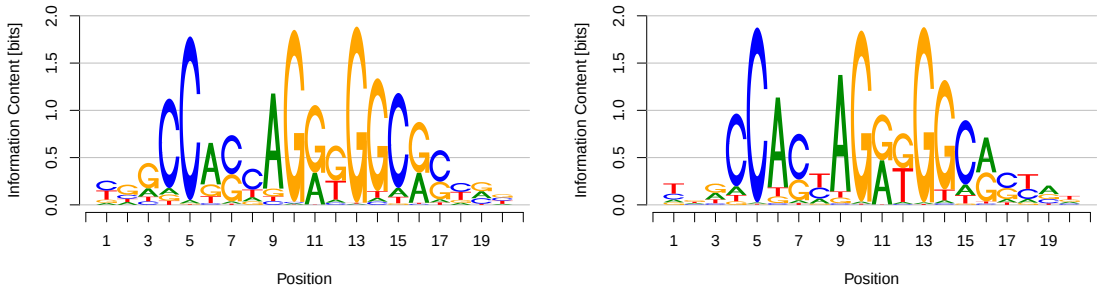


Figure 2.6: Sequence logos of the CTCF motifs from the cell lines H1-hESC (left) and HUVEC (right). The two sequence logos are highly similar in their conservation profile (height of stacks) and nucleotide preferences at the individual motif positions.

Sequence logos are used by researchers to interpret findings, document work, share knowledge, and present results. However, comparing multiple sequence logos by visual inspection is tricky, especially when the sequence motifs to compare are highly similar as in **Figure 2.6**. My colleagues and I address this problem in **Sections 3.3.1** and **6.1**.

2.4 Research objectives

The previous three sections introduced the reader to molecular biology, computer science, and bioinformatics. Six limitations in fields "data acquisition and data preparation," "*de–novo* motif discovery using *phylogenetic footprinting*," and "visualisation of sequence motifs" were shown, which my colleagues and I wish to address as follows.

First, we wish to improve knowledge sharing in the field of miRNA induced gene regulation and we wish to develop a new approach for the efficient storage of sequence related data with standard desktop hardware. Second, we wish to develop new approaches for *de–novo* motif discovery based on *phylogenetic footprinting*. Specifically, we propose an approach

based on *phylogenetic footprinting* to detect and correct the ubiquitous BA bias in ChIP-seq data. Further, we systematically study the influence of phylogenetic trees with different substitution probabilities on the classification performance of phylogenetic footprinting using synthetic and real data. Finally, we extend *phylogenetic footprinting* by the capability of taking into account intra-motif dependencies. Third, we wish to improve the comparative visualization of sequence motifs.

An important task in bioinformatics is the prediction of TFBSs in sequence data. This task typically starts with data acquisition and data preparation, e.g., sequence data are obtained from a ChIP-seq experiment. After transforming the data into an appropriate format, *de–novo* motif discovery is performed and putative TFBSs are predicted. Finally, the results are visualized and compared to related work. With this thesis, I wish to contribute to each of these three steps.

3 Context of publications

This chapter introduces the reader to the articles assembling this thesis. As indicated in **Section 2.4**, these articles can be divided into the three groups "Data acquisition and data preparation," "*De–novo* motif discovery using *phylogenetic footprinting*," and "Visualisation of sequence motifs." With all six publications and the one work-in-progress article my colleagues and I want to contribute to the understanding of the process of gene regulation and its evolution. **Figure 3.1** shows the applicability of the six peer reviewed publications and the one work-in-progress article and their relatedness to the process of gene regulation.

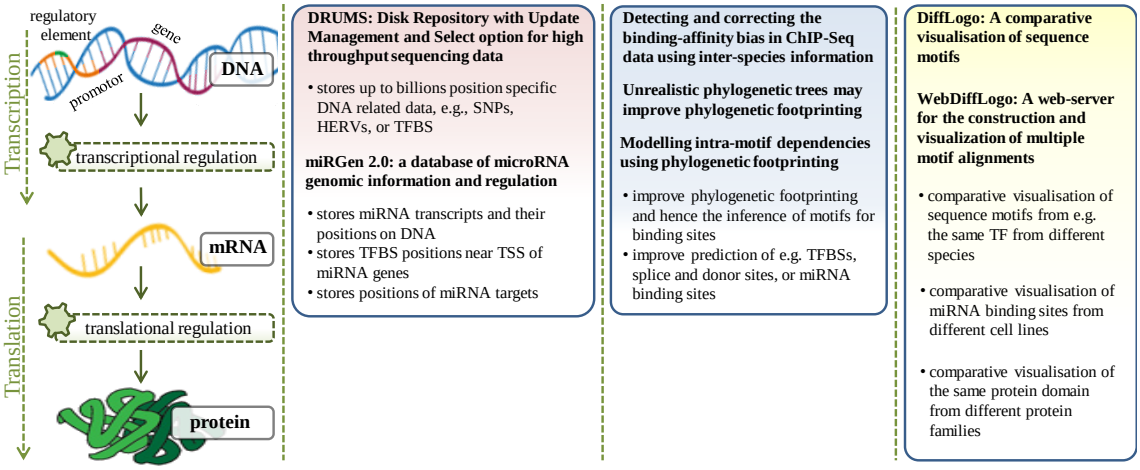


Figure 3.1: Articles of this thesis in context of gene expression. The first column summarizes the process of gene regulation. The second column (red) shows the two publications related to "Data acquisition and data preparation". The third column (blue) shows the three publications related to "*De–novo* motif discovery using *phylogenetic footprinting*". The fourth column (green) shows the publications and the one work-in-progress article related to "Visualisation of sequence motifs".

Data acquisition and data preparation are essential and preliminary tasks in all natural sciences. In the context of this thesis it would be impossible to perform *phylogenetic footprinting* without acquiring sequences from databases or similar sources and without aligning them in a preprocessing step. Further, the comparative visualization of sequence motifs using *DiffLogo* would be hardly possible without databases providing sequence motifs. Publications presented in this thesis related to this "Data acquisition and data preparation" are "miRGen 2.0: a database of microRNA genomic information and reg-

ulation" and "DRUMS: Disk Repository with Update Management and Select option for high throughput sequencing data". Regarding the process of gene regulation, *miRGen* supports researchers that are interested in miRNA regulation and miRNA function providing miRNA transcripts with target genes, SNPs, TFBSs in near distance, and prominent literature sources. Whereas *DRUMS* is applicable when dealing with hundred millions up to billions of DNA related information, like SNPs, TF binding site probabilities or human endogenous retrovirus (HERV) occurrences in the human genome.

De–novo motif discovery is an essential task in bioinformatics and a preliminary for understanding the process of gene regulation. *Phylogenetic footprinting* comprises approaches for *de–novo* motif discovery that take into account sequences of at least two phylogenetically related species. Publications presented in this thesis related to "*De–novo* motif discovery using *phylogenetic footprinting*" are "Detecting and correcting the binding-affinity bias in ChIP-seq data using inter-species information", "Unrealistic phylogenetic trees may improve phylogenetic footprinting", and "Combining phylogenetic footprinting with motif models incorporating intra-motif dependencies". All three proposed approaches may lead to an improved prediction of TFBSs and thus advance our understanding of transcriptional gene regulation and its evolution.

The visualization of results is an essential task in all sciences and it is needed to interpret findings, document work, share knowledge, and present results. Work related to "Visualisation of sequence motifs" in this thesis are the publication "DiffLogo: a comparative visualization of sequence motifs" and the work-in-progress article "WebDiffLogo: A webserver for the construction and visualization of multiple motif alignments".

The following subsections will introduce the reader into the publications and the one work-in-progress article comprising this work (**Figure 3.1**) and provide for each work a summary on the addressed objectives, the used methods, and the results. The full articles are presented in **Chapters 4**, **5**, and **6**.

3.1 Data acquisition and data preparation

The next two subsections provide a short summary of our publications "miRGen 2.0: a database of microRNA genomic information and regulation" and "DRUMS: Disk Repository with Update Management and Select option for high throughput sequencing data".

3.1.1 miRGen 2.0: a database of microRNA genomic information and regulation

The main objective of this work is to provide miRNA transcripts with related information to researchers of diverse disciplines who are interested in the regulation and function of miRNAs. Therefore, we collect miRNA transcripts from prominent literature sources and enrich these transcripts with information about TFBSs near the transcription start sites (TSS), miRNA expression profiles, and SNPs in miRNA hairpins.

Methods

In this work, we collect 812 human miRNA coding transcripts and 386 mouse miRNA coding transcripts from four literature sources. We identify for each miRNA primary transcript putative TFBSs in the region 5 kb upstream and 1 kb downstream of the TSS using *MatchTM* (Kel et al., 2003) and all vertebrate PWMs from Transfac 6.0 (Matys et al., 2003) minimizing the number of falsely predicted TFBSs. We provide for each predicted TFBS the matrix similarity score and the core similarity score calculated by *MatchTM*. We also identify miRNA expression profiles using the mammalian miRNA expression atlas (Ozsolak et al., 2008). We integrate data about SNPs located within the genomic positions corresponding to miRNA hairpins and TFBSs from the *UCSC* table browser (Karolchik et al., 2009). All data of the *miRGen* repository are stored using the relational database management system *MySQL*. **Figure 3.2** shows the relational schema of the database.

Results, discussion, and conclusions

Over 800 miRNA transcripts with TFBSs near their TSS, miRNA expression profiles, and SNPs in the miRNA hairpins are stored in the *miRGen* repository and are accessible via a user-friendly interface that allows searches for miRNAs and/or TFs of interest. The integration of the different information sources enables in-depth studies of miRNAs functions and contributes to the understanding of post-transcriptional gene regulation. Currently, *miRGen* is cited by more than 100 articles, specifically in the field of cancer research (Juan et al., n.d.; H.-D. Huang, 2012; Mar-Aguilar et al., 2016) and hence contributes to cancer diagnostics and therapeutics.

The TFBS annotations in *miRGen* from 2009 could be improved by using more sophisticated algorithms and motif models (instead of the PWM motif model) for the prediction of TFBSs. In addition, `http://www.factorbook.org/` (J. Wang et al., 2012) meanwhile provides extensive information for 167 TFs and their PWM representations for many exper-

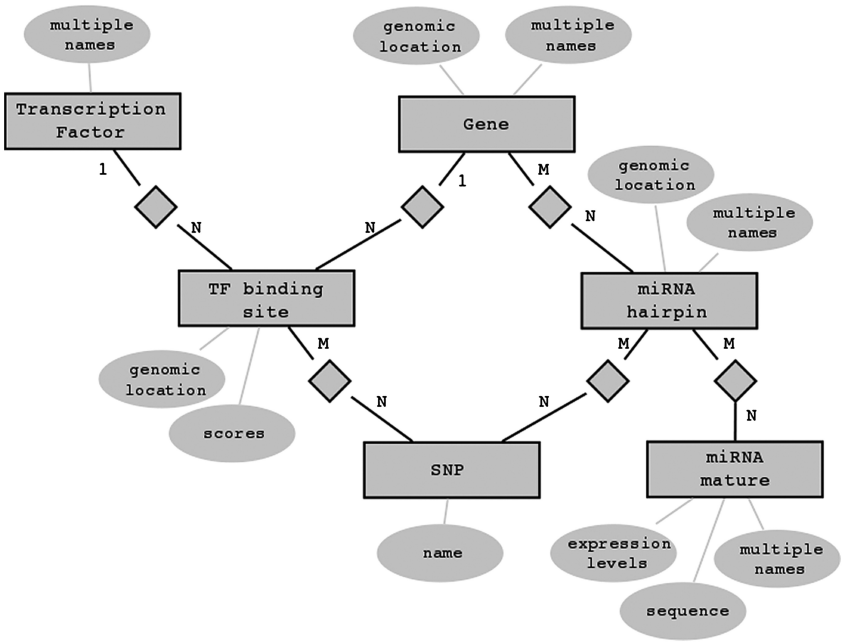


Figure 3.2: The *miRGen* database schema. TFs (top left) activate miRNA genes (top center) by binding to TFBSs (middle left). miRNA genes (top center) contain miRNA hairpins (middle right) that signify the genomic location of the mature miRNA-miRNA* duplex. miRNA hairpins are processed into mature miRNAs. Typically, one miRNA hairpin produces two mature miRNAs, but a mature miRNA can be produced by more than one hairpin from different miRNA genes. Both TFBSs and miRNA hairpins are genomic features that can contain SNPs. Mature miRNAs are associated with expression levels in various tissues and cell types.

iments from the ENCODE project (E. P. Consortium et al., 2004). Since many researchers use data from the ENCODE project for their research it would be of advantage if *miRGen* would use PWMs predicted on the very same data.

This work refers to version 2.0 of *miRGen*. The current version of *miRGen* is 3.0. *miRGen* is freely available at `http://www.microrna.gr/mirgen/`.

3.1.2 DRUMS: Disk Repository with Update Management and Select option for high throughput sequencing data

An important task of bioinformatics in the scope of computational biology projects is the efficient and well-organized data management. In fact often neglected, this issue becomes essential when dealing with many data sets that consist of millions or even billions of records. In addition, researchers from biology and biochemistry prefer to keep and analyze these data sets on a standard desktop machine for reasons of data privacy, limited computer skills, and convenience. Noble, 2009 describes how to organize data in computational biology projects and Wilson et al., 2014 describes best practices for scientific computing. In this work we tackle the problem of handling large data sets on a single standard desktop machine.

One kind of data extensively used in bioinformatics is position specific data related to DNA sequences. Examples are SNPs (*Single Nucleotide Polymorphism* 2012), transcription factor binding affinities, and probabilities (M. Bulyk, 2003; Nguyen et al., 2009), RNAseq data (Z. Wang et al., 2009; Malone et al., 2011), and mapping data from *BLAST* (M. Johnson et al., 2008). These data are essential for the understanding of biological and biochemical processes. We generalize this kind of data by the term position-specific DNA related data (psDrd).

Due to the rapid development of sequencing technologies and the ever-increasing number of tools and algorithms for analysing, manipulating, and combining psDrd, the data volume is growing exponentially. Thus, requesting data becomes challenging and expensive and is often tackled using specialised and/or distributed hardware. The objective of this work is the development of a data repository that is capable of storing billions of records of psDrd, performing fast and resource saving requests, and running on a single standard desktop hardware.

Methods

psDrd records have the following three characteristics that are important for finding or developing a suitable data repository. First, a psDrd record is representable by a key-value

pair, consisting of a unique key that defines a position on a sequence and a value that is associated to this sequence position. Second, all psDrd records of the same kind are storable using the same amount of memory. Third, researchers who work with psDrd are usually interested in all records near a certain sequence locus and the exact position of this locus is typically unknown.

By literature research we found the *DRUM* concept which was designed to store billions of URLs with meta data when crawling the world wide web (Lee et al., 2009). *DRUM* is already capable of storing large collections of key-value pairs by supporting fast bulk inserts without generating duplicate entries. Unfortunately, *DRUM* was not designed to request data in an efficient manner. We extended the *DRUM* concept by the capability of requesting a single record by key or a set of records in an interval between two keys. We developed the open-source *Java* library *DRUMS* meeting our requirements.

During the implementation of *DRUMS*, we focused on decoupling I/O-processes from memory processes to avoid blocking single components. We made extensively use of the prototype design pattern and the flyweight design pattern to reduce object instantiations. This relieves *Java's* object heap and hence dramatically reduces the number of runs of *Java's* garbage collector. **Figure 3.3** gives a high level overview of the insert process and the select process.

Results, discussion, and conclusions

We compared the performance of our implementation of *DRUMS* to the widely spread standard database *MySQL* on two data sets, considering database inserts, random lookups, and random range selects. The smaller data set contains SNPs for 251 accessions of the reference plant Arabidopsis thaliana. The larger data set represents a mapping of over 7000 HERV-fragments to the human genome which comprises more than 800 million records. In each test, *DRUMS* was considerably faster than *MySQL* by a factor of 2 up to a factor of 15456.

Based on this work, we added an additional feature to *DRUMS* which has not been evaluated systematically nor published yet. Namely, we added the capability of performing state dependent updates without rewriting or reorganising the files on disk. For example, to increment a counter in a traditional key-value store, first the counter is requested by key, then the fetched counter is incremented, and finally the incremented counter is written back as a new key-value pair. In contrast, *DRUMS* is capable of manipulating the corresponding data directly on disk resulting in a dramatic performance increase.

DRUMS is freely available at `http://mgledi.github.io/DRUMS/`.

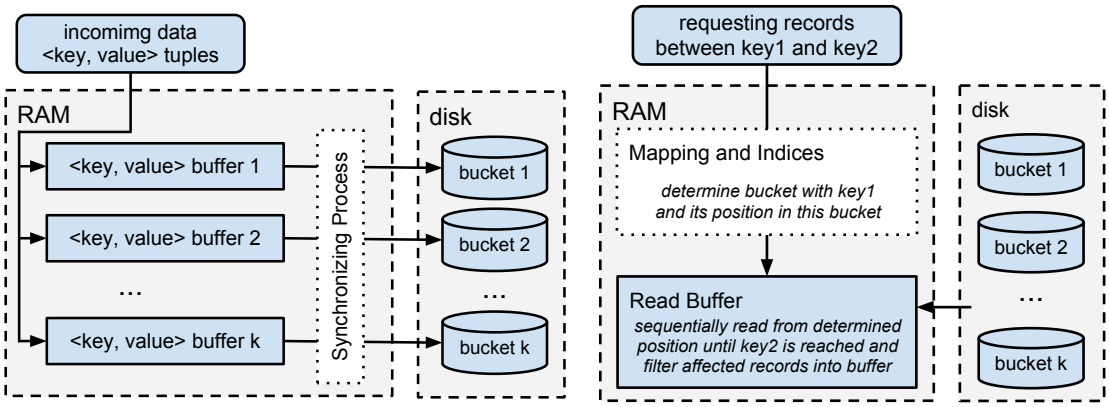


Figure 3.3: High level overview of the insert (left) and the select process (right) in *DRUMS*.
Insert process (left): Key-value pairs are sent to DRUMS. The incoming records are distributed between k buffers (memory buckets), based on their key. If a bucket B_i exceeds a predefined size or overall memory limitations are reached, a synchronisation process is instantiated.
Select process (right): A request is sent to DRUMS, typically providing a lower and an upper key. The following four steps are performed to get the requested records. 1) The bucket of interest is determined. 2) The correct chunk of the first requested record is identified, using a sparse index. 3) The position of the requested key-value pair is determined. 4) A sequential read is performed until the requested range is completely processed.

3.2 Predicting transcription factor binding sites using *phylogenetic footprinting*

The next three subsections provide a short summary of our publications "Detecting and correcting the binding-affinity bias in ChIP-seq data using inter-species information", "Unrealistic phylogenetic trees may improve phylogenetic footprinting", and "Combining phylogenetic footprinting with motif models incorporating intra-motif dependencies".

A common mathematical starting point for all three publications is the statistical model for a motif bearing alignment. The probability that the alignment X_n of length L_n is generated by the PFM as a motif bearing alignment is:

$$p(X_n|\theta) = \sum_{\ell_n=1}^{L_n-W+1} \frac{1}{L_n - W + 1} \prod_{u=1}^{L_n} \sum_{Y_n^u} p(Y_n^u|\ell_n, \theta) \prod_{o=1}^{O} p(X_n^{u,o}|Y_n^u, \ell_n, \theta) \tag{3.1}$$

where O denotes the number of species, W denotes the length of the motif, ℓ_n denotes the position of the motif, $X_n^{u,o}$ denotes the u-th symbol of the o-th sequence of the n-th align-

ment, and Y_n^u denotes the u-th symbol in the ancestral sequence. θ denotes the set of model parameters, namely the topology of the phylogenetic tree, the substitution probabilities, and the evolutionary model with its stationary probabilities for the flanking regions as well as for the binding site regions. I refer this formula in the following subsections.

3.2.1 Detecting and correcting the binding-affinity bias in ChIP-Seq data using inter-species information

The computational investigation of genomic regions containing TFBSs is a prerequisite for elucidating the process of gene regulation. ChIP-seq has become the major technology to uncover genomic regions containing TFBSs and countless approaches exist for predicting motifs from these genomic regions. It is known that ChIP-seq data and similar experimental data is contaminated with false positive genomic regions and there is evidence that there is an enrichment with high-affinity binding sites. Both factors potentially distort the results of *de novo* motif prediction which would affect all downstream analyses (Ross et al., 2013; Park et al., 2013; Teytelman et al., 2013; Elliott et al., 2015).

The contamination with false positive genomic regions leads to the *contamination bias* (Timothy L. Bailey, Krajewski, et al., 2013) and thus to the prediction of artificially softened motifs, whereas the enrichment of sequences with high-affinity binding sites leads to the *binding-affinity bias* (Håndstad et al., 2011) and thus to the prediction of artificially sharpened motifs. Most existing approaches for *de novo* motif prediction are capable of detecting and correcting the contamination bias and it has been shown that this increases the quality of motif prediction considerably (Timothy L. Bailey and Elkan, 1995; Wilbanks et al., 2010; Gomes et al., 2014).

The main objective of this work is to detect and correct the binding affinity bias and to improve *phylogenetic footprinting* by extending a traditional PFM that already takes into account the contamination bias by the capability to also take into account the BA bias.

Methods

To our knowledge, it is impossible to detect the BA bias based on sequence data from only one species, but detecting the BA bias appears to be possible using sequences from phylogenetically related species. The key idea is that mutations decrease the effect of BA bias in phylogenetically related species. Hence, the direct effect of the BA bias in the reference species is stronger than the indirect effect of the BA bias in phylogenetically related species. Under this assumption the information content of the predicted motif in the reference species should be higher than the information content of the predicted motifs in

phylogenetically related species. More specifically, the information content of the predicted motifs in the phylogenetically related species should decrease with the phylogenetic distance from the reference species. The detailed idea and a toy example can be found in the section "Using sequence information of phylogenetically related species to detect the binding-affinity bias" of the corresponding article (Nettling et al., 2016).

We investigate our hypothesis on 2132 sequence alignments comprising human ChIP-seq data of the five TFs CTCF, GABP, NRSF, SRF, and STAT1 with orthologous regions from the monkey, dog, cow, and horse by comparing the degree of information content in species-specific motifs. We propose a PFM capable of taking into account the contamination bias ($\mathcal{M}_{-}^{\mathrm{C}}$), the BA bias ($\mathcal{M}_{\mathrm{BA}}^{-}$), neither one or the other ($\mathcal{M}_{-}^{-}$), or both ($\mathcal{M}_{\mathrm{BA}}^{\mathrm{C}}$). We model the contamination bias using the popular zero or one occurrence of a binding site per sequence (ZOOPS) model, which is widely used for de-novo motif discovery (C. E. Lawrence et al., 1993; Redhead et al., 2007; Keilwagen et al., 2011; Agostini et al., 2014) and we model the effect of the BA bias using the Boltzmann distribution from thermodynamics (Maza et al., 1993). We transformed **Formula 3.1** in a way that the statistical model of a motif bearing alignment X_n of length L_n consisting of sequences from O species is defined as:

$$p(X_n|\theta) = \sum_{\ell_n=1}^{L_n-W+1} \frac{1}{L_n-W+1} \prod_{u=1}^{L_n} p(X_n^{u,1}|\ell_n,\theta) \tag{3.2}$$

$$\sum_{Y_n^u \in \mathcal{A}} p(Y_n^u|X_n^{u,1},\ell_n,\theta) \cdot \prod_{o=2}^{O} p(X_n^{u,o}|Y_n^u,\ell_n,\theta). \tag{3.3}$$

The inner factors of the sum are defined as follows:

$$p(X_n^{u,1}|\ell_n,\theta) = \begin{cases} \pi_0^a & \text{, if } u < \ell_n \text{ or } u \geq \ell_n + W \\ \frac{(\pi_{u-\ell_n+1}^a)^\beta}{\sum_{b\in\mathcal{A}}(\pi_{u-\ell_n+1}^b)^\beta} & \text{, if } \ell_n \leq u < \ell_n + W \end{cases} \tag{3.4}$$

$$p(Y_n = a|X_n^{u,1} = b,\ell_n,\theta) = \begin{cases} \gamma_1 \cdot \pi_0^a + (1-\gamma_1)\cdot\delta_{a=b} & \text{, if } u < \ell_n \text{ or } u \geq \ell_n + W \\ \gamma_1 \cdot \pi_{u-\ell_n+1}^a + (1-\gamma_1)\cdot\delta_{a=b} & \text{, if } \ell_n \leq u < \ell_n + W \end{cases} \tag{3.5}$$

$$p(X_n^{u,o} = a|Y_n = b,\ell_n,\theta) = \begin{cases} \gamma_o \cdot \pi_0^a + (1-\gamma_o)\cdot\delta_{a=b} & \text{, if } u < \ell_n \text{ or } u \geq \ell_n + W \\ \gamma_o \cdot \pi_{u-\ell_n+1}^a + (1-\gamma_o)\cdot\delta_{a=b} & \text{, if } \ell_n \leq u < \ell_n + W \end{cases} \tag{3.6}$$

where π_0^a denotes the probability of a base a in the background sequence, π_w^a denotes the probability of a base a in the motif sequence, γ_o denotes the substitution probability from the primordial species to species o.

β denotes the inverse temperature from the Boltzmann distribution and quantifies the degree of the BA bias in the reference species. We assume that a TF binds the binding site B with a probability proportional to $p(B|\pi)^{\beta-1}$. As B occurs in vivo with probability $p(B|\pi)$, it occurs in the set of immunoprecipitated sequences with a probability proportional to $p(B|\pi) \cdot p(B|\pi)^{\beta-1} = p(B|\pi)^{\beta}$. A value for β greater than one indicates that the ChIP-seq data set is affected by the binding-affinity bias, i.e., high-affinity binding sites are over-represented.

In **Supplementary Section 1** of the corresponding article and in **Section 7.1.1** of this thesis, the reader can find a detailed definition of the probabilistic model that is capable of taking into account both the contamination bias and the BA bias. In the corresponding article, we describe the PFM that is capable of taking into account the BA bias from the perspective of the data generating process.

We measure the classification performance of the four models $\mathcal{M}_{-}^{-}$, $\mathcal{M}_{\mathrm{BA}}^{-}$, $\mathcal{M}_{-}^{\mathrm{C}}$, and $\mathcal{M}_{\mathrm{BA}}^{\mathrm{C}}$ using 100 fold stratified repeated random sub-sampling validation. We calculate the information contents of the motifs predicted by the models taking into account the BA bias and the models neglecting the BA bias. We use *DiffLogo* to investigate differences in sequence motifs predicted by $\mathcal{M}_{\mathrm{BA}}^{-}$ and $\mathcal{M}_{\mathrm{BA}}^{\mathrm{C}}$ and the traditional motifs predicted by $\mathcal{M}_{-}^{-}$ and $\mathcal{M}_{-}^{\mathrm{C}}$.

Results, discussion, and conclusions

We found in case of all five TFs that the information contents of the human motifs are significantly higher than the information contents of the motifs from the monkey, dog, cow, and horse. We also found that the correction of the BA bias is possible using the proposed PFM leading to a more accurate inference of sequence motifs and to a more precise prediction of TFBSs. Interestingly, we found that the enrichment of ChIP-seq data with high-affinity binding sites causes a distortion of DNA binding motifs that is even stronger than the distortion caused by the contamination bias. The comparison of novel and traditional motifs showed small but noteworthy differences, suggesting that the refinement of traditional motifs from literature and databases might lead to the inference of novel binding sites, *cis*-regulatory modules, and gene-regulatory networks and may thus advance our attempt of understanding transcriptional gene regulation as a whole.

3.2.2 Unrealistic phylogenetic trees may improve *phylogenetic footprinting*

Two prerequisites for most *phylogenetic footprinting* algorithms are multiple sequence alignments (MSAs) of the DNA regions to analyse and phylogenetic trees, including substitution

probabilities attached to the branches. These phylogenetic trees are used to quantify the evolution of functional elements and their flanking DNA in the input MSAs. Hence, the choice of the phylogenetic trees and the substitution probabilities has a strong influence on the performance of *phylogenetic footprinting* and hence on the prediction of TFBSs (Kc et al., 2011). Typically, the phylogenetic trees used by nature to evolve the DNA regions of interest have been lost and are unknown. Estimating appropriate phylogenetic trees with appropriate substitution probabilities is hardly possible (Blanchette et al., 2003), so that the needed information is often simply taken from literature or guessed.

There are many articles that state that *phylogenetic footprinting* improves motif prediction but none of them investigates the influence of different phylogenetic trees on classification performance and motif prediction on non-synthetic data (Moses et al., 2004; Gertz et al., 2006; Clark et al., 2007; Hardison et al., 2012). Thus, the main objective of this work is to study systematically the influence of the phylogenetic trees on the performance of *phylogenetic footprinting*.

Methods

To systematically investigate the influence of phylogenetic trees on the performance of *phylogenetic footprinting* we made the following simplification. The PFM uses a star topology instead of a more complex phylogenetic tree with all branches having the same length, i.e., the substitution probability γ is the same for all species. With this simplification it is now possible to investigate the performance of a PFM as function of the substitution probability γ, where small γ encode closely phylogenetic relations and large γ encode loosely phylogenetic relations. The statistical model of a motif bearing alignment looks different to **Formula 3.1** but is the same. We extracted the parameter γ from the set of parameter set θ. The probability that the alignment X_n of length L_n consisting of sequences from O observed species can be calculated as follows:

$$p(X_n|\gamma,\theta) = \sum_{\ell_n=1}^{L_n-W+1} \frac{1}{L_n-W+1} \prod_{u=1}^{L_n} \sum_{Y_n^u} p(Y_n^u|\ell_n,\gamma,\theta) \prod_{o=1}^{O} p(X_n^{u,o}|Y_n^u,\ell_n,\gamma,\theta) \quad (3.7)$$

The inner factors are defined as follows:

$$p(Y_n^u|\ell_n,\gamma,\theta) = \begin{cases} \pi_0^a & \text{if } u < \ell_n \text{ or } u \geq \ell_n + W \\ \pi_{u-\ell_n+1}^a & \text{if } \ell_n \leq u < \ell_n + W \end{cases} \quad (3.8)$$

$$p(X_n^{u,k}|Y_n^u,\ell_n,\gamma,\theta) = \begin{cases} \gamma \times \pi_0^a + (1-\gamma)\delta_{a=b} & \text{if } u < \ell_n \text{ or } u \geq \ell_n + W \\ \gamma \times \pi_{u-\ell_n+1}^a + (1-\gamma)\delta_{a=b} & \text{if } \ell_n \leq u < \ell_n + W \end{cases} \quad (3.9)$$

where π_0^a denotes the probability of a base a in the background sequence, π_w^a denotes the probability of a base a in the motif sequence, γ_k denotes the substitution probability from the primordial species to species k. The complete statistical model and all parameters are explained in **Methods 1** of the corresponding article.

We investigate the classification performance and the likelihood of the PFM for $\gamma = \{0.05, 0.1, \ldots, 1.0\}$ on human ChIP-seq data of the five TFs CTCF, GABP, NRSF, SRF, and STAT1 enriched with orthologous regions from the monkey, dog, cow, and horse as well as on synthetic data generated using a PFM with $\gamma = 0.2$. We further compare the classification performance of the three PFMs using a tree from literature (Arnold et al., 2012) ($\mathcal{M}_{lit}^{tree}$), a star topology with the maximum likelihood estimated γ ($\mathcal{M}_{ML}^{star}$), and a star topology with $\gamma = 1$ ($\mathcal{M}_{\gamma=1.0}^{star}$).

Results, discussion, and conclusions

When studying the likelihood, we found that on synthetic data the best likelihood is achieved when using the same phylogenetic tree for learning as for data generation. We also observed that on organic data the best likelihood is achieved when using realistic phylogenetic trees indicating that we are capable of identifying reasonable substitution probabilities for synthetic and for real data using the maximum-likelihood principle.

When investigating the classification performance, we found that on synthetic data the best classification performance is achieved when using the same phylogenetic tree for learning as for data generation. In contrast, we found that on organic data unrealistic phylogenetic trees often lead to more accurate predictions of transcription factor binding sites than realistic phylogenetic trees. We also observed that $\mathcal{M}_{\gamma=1.0}^{star}$ significantly outperforms $\mathcal{M}_{ML}^{star}$ and $\mathcal{M}_{lit}^{tree}$. With other words, choosing unrealistic model assumptions with *phylogenetic footprinting* – namely using a star topology with unrealistic large substitution probabilities – may yield higher classification performances than using realistic phylogenetic trees with more realistic substitution probabilities.

Although we have no concrete explanation for this observation, we speculate that evolutionary effects like heterogeneous and heterotachious substitution probabilities among different DNA positions are violating model assumptions like the assumption of time reversibility. Further, these effects might already have lead to the construction of incorrect or at least partially erroneous MSAs. PFMs using a star topology with substitution probabilities of $\gamma = 1$ seem to be more robust toward these effects than PFMs using realistic phylogenetic trees with realistic substitution probabilities. Hence, we need to give the strange but practical recommendation to use PFMs based on these unrealistic model assumptions until there are more appropriate PFMs that take into account heterogeneity and heterotachy as well as putative misalignments in the input MSAs.

3.2.3 Combining *phylogenetic footprinting* with motif models incorporating intra-motif dependencies

As stated repeatedly in this work, de-novo motif discovery is a challenging task in bioinformatics and many different approaches exist for solving this task. These approaches can be divided in two groups.

The first group comprises approaches using sequences of only one species, which we refer to as one-species approaches. Within this group, a variety of statistical models are used for the binding of TFs to their TFBSs, ranging from the simple PWM model, neglecting intra-motif dependencies, to more complex models, taking into-account intra-motif dependencies (Grau, Posch, et al., 2013; I. Kulakovskiy et al., 2013; Ma et al., 2014; Alipanahi et al., 2015; Siebert et al., 2016).

The second group comprises approaches using sequences of at least two phylogenetic related species, which is known as *phylogenetic footprinting*. Within this group, statistical models are used that are capable of modeling the binding of TFs to their TFBSs and their evolution simultaneously (Blanchette et al., 2003; Sinha et al., 2004; Moses et al., 2004; Siddharthan et al., 2005; Neph et al., 2006; Newberg et al., 2007; Siddharthan, 2008; Arnold et al., 2012).

It has been shown that more complex motif models taking into account intra-motif dependencies outperform simpler motif models like the PWM model and that models that take into account phylogenetic dependencies also outperform the PWM model (M. L. Bulyk et al., 2002; Salama et al., 2010; Eggeling et al., 2015). One-species approaches neglect phylogenetic information, whereas *phylogenetic footprinting*, which incorporates this information, neglects intra-motif dependencies.

The main objective of this work is to improve *phylogenetic footprinting* by taking into account base dependencies, i.e., developing an approach for de-novo motif discovery that takes into account both phylogenetic dependencies and base dependencies simultaneously.

Methods

We extend a PFM model based on the Felsenstein evolutionary model (Felsenstein, 1981) by the capability of taking into account base dependencies resulting in a model that is capable of taking into account base dependencies and phylogenetic dependencies simultaneously. We use **Formula 3.1** as starting point and split up the product $\prod_{u=1}^{L_n}$ that takes into account the whole alignment into three products, namely the region left of the motif, the

motif region, and the region right of the motif. The statistical model of a motif bearing alignment X_n of length L_n consisting of sequences from O species is defined as:

$$p(X_n, \theta) = \sum_{\ell_n=1}^{L_n-W+1} \frac{1}{L_n - W + 1} \prod_{o=1}^{O} p(X_n^{i(\ell_n),o}|\ell_n, \theta) \cdot p(X_n^{m(\ell_n),o}|\ell_n, \theta) \cdot p(X_n^{e(\ell_n),o}|\ell_n, \theta) \tag{3.10}$$

The inner factors are defined as follows:

$$p(X_n^{i(\ell_n),o}|\ell_n, \theta) = \prod_{u \in \{1,\ldots,\ell_n-1\}} \pi_0^{a,\zeta} \quad \textit{(left flanking region)} \tag{3.11}$$

$$p(X_n^{m(\ell_n),o}|\ell_n, \theta) = \prod_{u \in \{\ell_n,\ldots,\ell_n+W-1\}} \pi_{u-\ell_n+1}^{a,\zeta} \quad \textit{(motif region)} \tag{3.12}$$

$$p(X_n^{e(\ell_n),o}|\ell_n, \theta) = \prod_{u \in \{\ell_n+W,\ldots,L_n\}} \pi_0^{a,\zeta} \quad \textit{(right flanking region)} \tag{3.13}$$

where the probability of a base a in the background sequence provided that its predecessors are in joint state ζ is given by the parameter $\pi_0^{a,\zeta}$ and the probability of a base a in the motif sequence provided that its predecessors are in joint state ζ is given by the parameter $\pi_w^{a,\zeta}$. The complete statistical model and all parameters are explained in **Methods 2** of the corresponding article.

We study the proposed PFMs on datasets based on ChIP-seq data of 35 TFs. First, we measure the degree of intra-motif dependencies captured by the proposed PFMs by computing the position–wise mutual information (mutual information). We call the resulting vector of mutual information values mutual information profile. For each of the 35 TFs, we compute the mutual information profiles of orders 1 and 2 from the motifs obtained by the PFMs of order 2. Moreover, we study for each TF the similarity of the species–specific motifs using *DiffLogo*and the similarity of the species–specific mutual information profiles using statistical tests.

Second, we study the classification performance of the PFMs of orders 0, 1, and 2 using 25–fold stratified repeated random sub-sampling validation. We calculate and show for each TF the relative increase of the PFMs of orders 1 and 2 relative to the PFM of order 0 Moreover, we compare the classification performance of phylogenetic footprinting and one-species approaches when neglecting and taking into account base dependencies of order 2.

Results, discussion, and conclusions

First, we found for the studied TFs statistically significant intra-motif dependencies between neighboring bases at all positions and we found even stronger intra-motif dependencies between dimers and their neighboring bases at all positions. We excluded the possibility that the captured intra-motif dependencies are an artifact resulting from a mixture of different species–specific motifs.

Second, we found that modeling base dependencies of order 1 improves *phylogenetic footprinting* for 31 TFs and we found that modeling base dependencies of order 2 improves *phylogenetic footprinting* for all 35 TFs and always outperforms modeling base dependencies of order 1. By comparing the classification performances of the four cases of one-species approaches and *phylogenetic footprinting* when neglecting and taking into account base dependencies, we found that taking into account both phylogenetic dependencies and base dependencies outperforms the other three approaches in 31 of the 35 TFs.

These findings suggest that combining *phylogenetic footprinting* with motif models incorporating intra-motif dependencies may lead to an improved prediction of TFBSs and thus advance our understanding of transcriptional gene regulation and its evolution.

3.3 Visualization of sequence motifs

The next two subsections provide a short summary of the publication "DiffLogo: a comparative visualization of sequence motifs" and the work-in-progress article "WebDiffLogo: A web-server for the construction and visualization of multiple motif alignments".

3.3.1 DiffLogo: A comparative visualization of sequence motifs

An important task in bioinformatics is the visualization of results from the analysis of biological data. In the field of *de-novo* motif discovery and TFBSs prediction, sequence motifs are used to represent functional regions of biological sequences, e.g., TFBSs, splice sites in pre-mRNAs, miRNA binding sites, or phosphorylation sites of proteins. Sequence logos are the *de facto* standard for the visualization of these sequence motifs and are essential for researchers to interpret findings, document work, share knowledge, and present results (Schneider et al., 1990).

Due to the increasing number of datasets and due to the increasing number of approaches for *de-novo* motif discovery, the research focus has shifted from inferring only "the" sequence motif towards comparative analyses to study the reasons for, e.g., the differential binding of TFBS under different conditions. Sequence logos are not suited for the discovery of the,

sometimes subtle, differences between, e.g., cell type - specific sequence motifs resulting from the differential binding of the TF of interest.

Initial approaches for comparative visualization of sequence motifs can be found in *iceLogo* (Colaert et al., 2009; Maddelein et al., 2015), *MotifStack* (Jianhong Ou, 2014), *STAMP* (Mahony et al., 2007), and *Two Sample Logo* (Vacic et al., 2006). None of them allows a configurable and comparative visualization of multiple sequence motifs. Table 1 in "DiffLogo: a comparative visualization of sequence motifs" shows an comparative overview of the main features of each tool. The objective of this work is to develop an easy to use and configurable tool for the comparative visualization of multiple sequence motifs.

Methods

Inspired by the intuitive *sequence logo* approach, we propose the *difference logo* to present differences between two sequence motifs. A *difference logo* depicts position-wise differences between two motifs of length L by L symbol stacks. The height of a stack is proportional to the degree of symbol distribution dissimilarity and the height of a symbol is proportional to the degree of differential symbol abundance. In case of $N > 2$ motifs, we take into account all $N \times (N-1)$ pair-wise motif comparisons and arrange the resulting *difference logos* in a $N \times N$ grid with one row and one column for each motif. Similar motifs are placed in nearby rows and columns.

Since the software environment R already has a large community among researchers from natural sciences, we implement this idea using R and make it publicly available in the R package *DiffLogo*. By default, *DiffLogo* uses the Jensen-Shannon divergence to calculate symbol distribution differences depicted by the height of the symbol stack at position ℓ:

$$H_\ell = \frac{1}{2} \sum_{a \in \mathcal{A}} p_{\ell,a} \log_2 \frac{p_{\ell,a}}{m_{\ell,a}} + \frac{1}{2} \sum_{a \in \mathcal{A}} q_{\ell,a} \log_2 \frac{q_{\ell,a}}{m_{\ell,a}},$$

where $m_{\ell,a} = \frac{p_{\ell,a}+q_{\ell,a}}{2}$.

The height of a symbol in stack ℓ is by default determined by the probability difference normalized by the sum of absolute probability differences at position ℓ:

$$H_{\ell,a} = H_\ell \cdot \begin{cases} \frac{p_{\ell,a}-q_{\ell,a}}{\sum_{a' \in \mathcal{A}} |p_{\ell,a'}-q_{\ell,a'}|} & \text{if } p_\ell \neq q_\ell \\ 0 & \text{otherwise.} \end{cases}$$

DiffLogo orders sequence motifs using hierarchical clustering and optimal leaf ordering to ensure that similar motifs are close to each other in the $N \times N$ grid of *difference logos*. The viewer is able to overlook the overall motif differences by the background color of each *difference logo* and a leaf-ordered cluster tree on top of the grid.

Results, discussion, and conclusions

We developed the *R* package *DiffLogo* for the visualization of differences between various types of sequence motifs. We demonstrated the utility of *DiffLogo* using binding motifs of the human insulator CTCF from different cell types and successfully reproduced findings from literature. In addition, we applied *DiffLogo* to E-box motifs of three basic helix-loop-helix transcription factors and to the F-Box binding domain from three different species groups revealing noteworthy motif differences.

Using *DiffLogo*, it is easily possible to compare motifs from different sources. Hence, *DiffLogo* facilitates decision making, knowledge sharing, and the presentation of results. The *DiffLogo* package comprises example data, example code, and further documentation and is freely available at *Bioconductor*[1]. In 2016, *DiffLogo* was downloaded more than 100 times per month and more than 1000 times in total.

3.3.2 WebDiffLogo: A web-server for the construction and visualization of multiple motif alignments

In the previous work (Nettling, Treutler, Grau, et al., 2015), we presented *DiffLogo*, an *R* package developed for the comparative visualization of sequence motifs. We think that *DiffLogo* is already easy to use, but we know that not all researches have access to hardware with *R* and *DiffLogo* installed and that not all researches have enough technical background or the time to use *R* and *DiffLogo* without high effort. Another hindering preliminary is that all input sequence motifs must have the same length and the same orientation in case of TFBSs to get meaningful *difference logos*. We experienced that this is often hard to accomplish, especially for users that do not have any background in bioinformatics.

The objective of this work is to make *DiffLogo* usable to researchers that are less experienced with the *R* programming language and to researchers without any background in bioinformatics. First, we extend *DiffLogo* by the capability to align sequence motifs. Second, we integrate *DiffLogo* into the intuitive to use web-server *WebDiffLogo* accessible via `http://difflogo.com`.

[1] `http://bioconductor.org/packages/release/bioc/html/DiffLogo.html`

Methods

First, the multiple motif alignment is computed by adjusting the relative shifts and relative orientations of the single sequence motifs based on a heuristic algorithm using the UPGMA algorithm and an extension of the sum-of-pairs score from symbols to conditional probability distributions (Sokal, 1958; Wheeler et al., 2007). We adapted the visualization of difference logos indicating unaligned flanking regions with a gray background. Further, all sequence logos and difference logos in a table of difference logos are shown aligned. Thus, a visual inspection of sequence logos and difference logos can be easily accomplished.

Second, we integrated *DiffLogo* into the web-server `http://difflogo.com`. Frontend and backend are fully implemented using the *Javascript* library *ReactJS*[1]. The front-end is designed as a single page application that permanently communicates with the web-server to e.g., validate files or generate sequence logos. This gives the user the feeling of a desktop application.

The source code of the web-server is publicly available at `https://github.com/mgledi/DiffLogoUI`.

Results, discussion, and conclusions

WebDiffLogo allows building difference logos of two sequence motifs and tables of difference logos of more than two sequence motifs via a user friendly single page application. The user starts by uploading the input set of sequence motifs or the input set of sets of aligned sequences in one of several common formats[2]. *WebDiffLogo* converts collections of aligned sequences to sequence motifs automatically, and the user is then allowed to select sequence motifs for the subsequent comparison. Sequence motifs of different length and different orientation will be automatically aligned. *WebDiffLogo* finally compares the aligned motifs in a position-wise manner and visualizes the over- and under-represented symbols as difference logos. The visualization is shown in the browser window. The output data are kept for 24 hours and can be downloaded by the user as PNG files as well as publication-ready vector graphics files. The user can also download and adapt the *R* code that generated the results. Figure 3.4 shows an example difference logo and the sequence logos of two CTCF motifs differing length and strand orientation.

[1] `http://github.com/facebook/react`
[2] `https://github.com/mgledi/DiffLogoUI/wiki/Supported-file-formats`

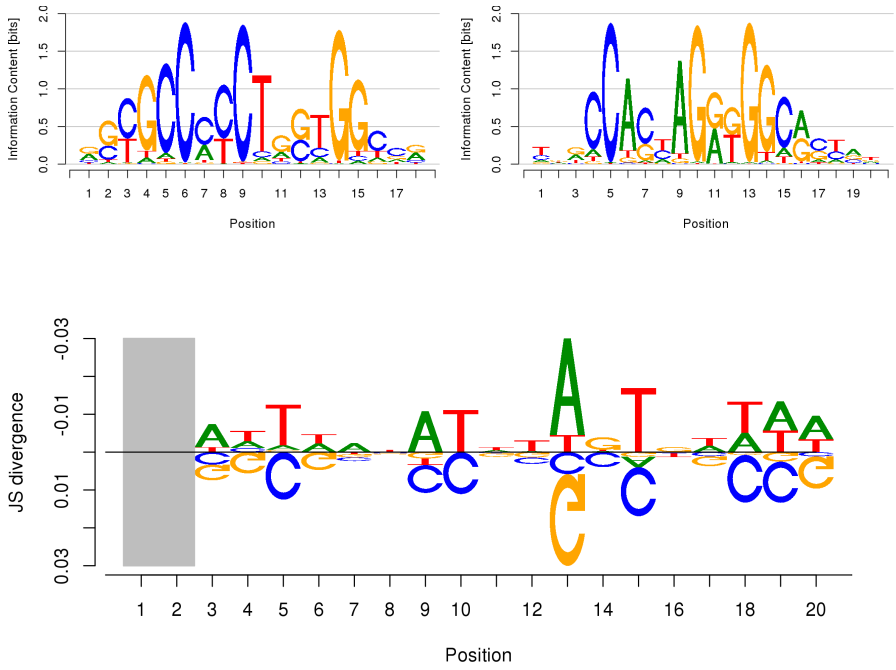


Figure 3.4: Difference logo (bottom) and sequence logos of CTCF motifs from cell lines H1-hESC (top left) and HUVEC (top right). The H1-hESC motif is two bases shorter than the HUVEC motif. The two motifs also differ in their strand orientation. The resulting difference logo depicts the small differences of the aligned motifs. Unaligned regions are indicated with gray background.

3.4 Conclusions and outlook

In this thesis, my colleagues and I have addressed six limitations in three related fields.

First, we proposed *miRGen* and *DRUMS*, two approaches to improve "data acquisition and data preparation." Specifically, by providing access to over 800 miRNA transcripts enriched with information about TFBSs near their TSS, with miRNA expression profiles, and with SNPs, *miRGen* improves the insights into the involvement of miRNAs in gene regulation and thus contributes to cancer diagnostics and therapeutics. Further, *DRUMS* is a key-value store optimized to handle psDrd running on standard desktop hardware. *DRUMS* is considerably faster than *MySQL* by a factor of 2 up to a factor of 15456 regarding this kind of data.

In comparison to many other NoSQL databases, neither is *DRUMS* horizontally scalable nor does it support redundancy. It would be an interesting follow-up project to investigate if the *DRUMS* concept could be extended by those capabilities.

Second, we proposed three approaches to improve "*de–novo* motif discovery using *phylogenetic footprinting*." Specifically, we found that it is possible to detect and correct the BA bias using inter–species information and that taking into account this bias leads to a more precise prediction of TFBSs using *phylogenetic footprinting* on ChIP-seq data. Further, we found that *phylogenetic footprinting* using a star topology with unrealistic high substitution probabilities seem to be more robust toward violation of model assumptions caused by evolutionary effects like heterogeneous and heterotachious substitution probabilities. Finally, we found that combining *phylogenetic footprinting* with motif models incorporating intra–motif dependencies lead to an improved prediction of TFBSs. Each of these findings advance our attempt of understanding transcriptional gene regulation as a whole.

Regarding "*De–novo* motif discovery using *phylogenetic footprinting*," I propose the following future works. It would be interesting to investigate more complex evolutionary models like the HKY model or the GTR model with the *phylogenetic footprinting* approaches proposed in this thesis . Further, it seems promising to combine the *phylogenetic footprinting* approaches proposed in this thesis to one approach that is capable of detecting and correcting the BA bias and to model intra-motif dependencies. Finally, a combination of *phylogenetic footprinting* with more complex motif models like parsimonious Markov models or Bayesian Markov models could improve *de–novo* motif discovery based on *phylogenetic footprinting* (Eggeling et al., 2015; Siebert et al., 2016).

Third, we proposed *DiffLogo* to improve "visualisation of sequence motifs." *DiffLogo* is an *R* package publicly available via Bioconductor[1] or GitHub[2], developed for the comparative

[1] `http://bioconductor.org/packages/release/bioc/html/DiffLogo.html`
[2] `http://github.com/mgledi/DiffLogo`

visualization of sequence motifs. *DiffLogo* was downloaded more than 100 times per month and more than 1000 times in total in 2016. To make *DiffLogo* applicable to a broader usership, we integrated the *R* package into the easy-to-use web-server *WebDiffLogo*[1]. *DiffLogo* and *WebDiffLogo* facilitate decision making, knowledge sharing, and the presentation of results.

An interesting extension of *DiffLogo* could be the higher order comparative visualization of sequence motifs, i.e., the comparative visualization of intra–motif dependencies. An interesting follow-up of *WebDiffLogo* could be a web-server that allows the investigation of sequence motifs with several existing tools like *IceLogo*, *motifStack*, or *Two Sample Logo* (Colaert et al., 2009; Jianhong Ou, 2014; Vacic et al., 2006). Additionally, the investigation of sequence motifs could be dramatically improved by allowing user interactions with the results, e.g., investigating only a few motif positions of interest.

[1] `http://difflogo.com`

Glossary

AUC area under receiver operating characteristics curve.

BA bias binding affinity bias.

bp base pair.

ChIP-seq ChIP-sequencing.

DNA Deoxyribonucleic acid.

DRUM Disk Repository with Update Management.

DRUMS Disk Repository with Update Management and Select option.

EM Expectation Maximization.

F81 evolutionary model Felsenstein 81.

HERV human endogenous retrovirus.

hMM(k) homogeneous Markov model of order k.

iMM(k) inhomogeneous Markov model of order k.

KLD Kullback—Leibler divergence.

miRNA microRNA.

mRNA messenger RNA.

MSA multiple sequence alignment.

mutual information mutual information.

NoSQL not only SQL.

PFM phylogenetic footprinting model.

PR precision recall.

psDrd position-specific DNA related data.

PWM position weight matrix.

RISC RNA-Induced Silencing Complex.

RNA Ribonucleic acid.

ROC receiver operating characteristics.

SNP single nucleotide polymorphism.

SQL structured query language.

TF transcription factor.

TFBS transcription factor binding site.

TSS transcription start site.

UTR untranslated region.

ZOOPS zero or one occurrence of a binding site per sequence.

References

Agostini, Federico, Davide Cirillo, Riccardo D Ponti, and Gian G Tartaglia (2014). SeAMotE: a method for high-throughput motif discovery in nucleic acid sequences. *BMC genomics*, 15 (1), p. 925.

Alexiou, Panagiotis, Thanasis Vergoulis, Martin Gleditzsch, George Prekas, Theodore Dalamagas, Molly Megraw, Ivo Grosse, Timos Sellis, and Artemis G Hatzigeorgiou (2009). miRGen 2.0: a database of microRNA genomic information and regulation. *Nucleic acids research*, gkp888.

Alipanahi, Babak, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33 (8), pp. 831–838.

Arnold, Phil, Ionas Erb, Mikhail Pachkov, Nacho Molina, and Erik van Nimwegen (2012). MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics*, 28 (4), pp. 487–494.

Bailey, Timothy L. and Charles Elkan (1995). "The value of prior knowledge in discovering motifs with MEME." In: *ISMB. International Conference on Intelligent Systems for Molecular Biology*. Vol. 3, pp. 21–29.

Bailey, Timothy L., Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS computational biology*, 9 (11).

Bailey, Timothy, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol*, 9 (11), e1003326.

Bembom, Oliver. *seqLogo: Sequence logos for DNA sequence alignments*. R package version 1.28.0. Division of Biostatistics, University of California, Berkeley.

Blanchette, Mathieu and Martin Tompa (2003). FootPrinter: a program designed for phylogenetic footprinting. *Nucleic acids research*, 31 (13), pp. 3840–3842.

Bourgon, Richard, Robert Gentleman, and Wolfgang Huber (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107 (21), pp. 9546–9551.

Bulyk, Martha (2003). Computational prediction of transcription-factor binding site locations. *Genome Biology*, 5 (1), pp. 201+.

Bulyk, Martha L, Philip LF Johnson, and George M Church (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research*, 30 (5), pp. 1255–1261.

Calin, George A and Carlo M Croce (2006). MicroRNA signatures in human cancers. *Nature Reviews Cancer*, 6 (11), pp. 857–866.

Carrington, James C and Victor Ambros (2003). Role of microRNAs in plant and animal development. *Science*, 301 (5631), pp. 336–338.

Chen, Kevin and Nikolaus Rajewsky (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics*, 8 (2), pp. 93–103.

Clark, Andrew G, Michael B Eisen, Douglas R Smith, Casey M Bergman, Brian Oliver, Therese A Markow, Thomas C Kaufman, Manolis Kellis, William Gelbart, Venky N Iyer, et al. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450 (7167), pp. 203–218.

Codd, Edgar F (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13 (6), pp. 377–387.

Colaert, Niklaas, Kenny Helsens, Lennart Martens, Joel Vandekerckhove, and Kris Gevaert (2009). Improved visualization of protein consensus sequences by iceLogo. *Nature methods*, 6 (11), pp. 786–787.

Consortium, 1000 Genomes Project et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491 (7422), pp. 56–65.

Consortium, ENCODE Project et al. (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306 (5696), pp. 636–640.

— (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489 (7414), pp. 57–74.

Crick, Francis et al. (1970). Central dogma of molecular biology. *Nature*, 227 (5258), pp. 561–563.

Darwin, Charles (1859). *On the origin of species*.

D'haeseleer, Patrik (2006). How does DNA sequence motif discovery work? *Nature biotechnology*, 24 (8), pp. 959–961.

Dolinoy, Dana C, Jennifer R Weidman, and Randy L Jirtle (2007). Epigenetic gene regulation: linking early developmental environment to adult disease. *Reproductive Toxicology*, 23 (3), pp. 297–307.

Dragland, Åse (2013). *Big Data, for better or worse: 90% of world's data generated over last two years*. URL: `https://www.sciencedaily.com/releases/2013/05/130522085217.htm` (visited on 12/10/2016).

Durinck, Steffen, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21 (16), pp. 3439–3440.

Eggeling, Ralf, Teemu Roos, Petri Myllymäki, and Ivo Grosse (2015). Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC bioinformatics*, 16 (1), p. 375.

Elliott, Julian H., Jeremy Grimshaw, Russ Altman, Lisa Bero, Steven N. Goodman, David Henry, Malcolm Macleod, David Tovey, Peter Tugwell, Howard White, and Ida Sim (2015). Informatics: Make sense of health data. *Nature*, 527, pp. 31–32.

Farazi, Thalia A, Jessica I Spitzer, Pavel Morozov, and Thomas Tuschl (2011). miRNAs in human cancer. *The Journal of pathology*, 223 (2), pp. 102–115.

Felsenstein, Jospeh (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17 (6), pp. 368–376.

Fougerolles, Antonin de, Hans-Peter Vornlocher, John Maraganore, and Judy Lieberman (2007). Interfering with disease: a progress report on siRNA-based therapeutics. *Nature reviews Drug discovery*, 6 (6), pp. 443–453.

Fredslund, Jakob (2006). PHY·FI: fast and easy online creation and manipulation of phylogeny color figures. *BMC bioinformatics*, 7 (1), p. 1.

Furey, Terrence S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics*, 13 (12), pp. 840–852.

Gantz, John and David Reinsel (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007, pp. 1–16.

Gentleman, Robert C, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5 (10), R80.

Gertz, Jason, Justin C. Fay, and Barak A. Cohen (2006). Phylogeny based discovery of regulatory elements. *BMC Bioinformatics*, 7, p. 266.

Gomes, Antonio LC, Thomas Abeel, Matthew Peterson, Elham Azizi, Anna Lyubetskaya, Luís Carvalho, and James Galagan (2014). Decoding ChIP-seq with a double-binding signal refines binding peaks to single-nucleotides and predicts cooperative interaction. *Genome research*, 24 (10), pp. 1686–1697.

Gosling, James, Bill Joy, Guy L Steele, Gilad Bracha, and Alex Buckley (2014). *The Java Language Specification*. Pearson Education.

Grau, Jan, Jens Keilwagen, André Gohr, Berit Haldemann, Stefan Posch, and Ivo Grosse (2012). Jstacs: A Java framework for statistical analysis and classification of biological sequences. *Journal of Machine Learning Research*, 13 (Jun), pp. 1967–1971.

Grau, Jan, Stefan Posch, Ivo Grosse, and Jens Keilwagen (2013). A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic acids research*, gkt831.

Håndstad, Tony, Morten B. Rye, Finn Drabløs, and Pål Sætrom (2011). A ChIP-Seq Benchmark Shows That Sequence Conservation Mainly Improves Detection of Strong Transcription Factor Binding Sites. *PLoS ONE*, 6 (4), e18430+.

Hardison, Ross C and James Taylor (2012). Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics*, 13 (7), pp. 469–483.

He, Lin and Gregory J Hannon (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, 5 (7), pp. 522–531.

Huang, Hsien-Da (2012). MicroRNA Research in Cancer Biology: Databases and Tools. In: *Systems Biology: Applications in Cancer-Related Research*, pp. 209–224.

Huber, Wolfgang, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*, 12 (2), pp. 115–121.

Jianhong Ou, Lihua Julie Zhu (2014). *motifStack: Plot stacked logos for single or multiple DNA, RNA and amino acid sequence*. URL: `http://www.bioconductor.org/packages/release/bioc/html/motifStack.html`.

Johnson, David S, Ali Mortazavi, Richard M Myers, and Barbara Wold (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316 (5830), pp. 1497–1502.

Johnson, Mark, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott D. McGinnis, and Thomas L. Madden (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, 36 (Web-Server-Issue), pp. 5–9.

Juan, Hsueh-Fen, Hsuan-Cheng Huang, Feng-Sheng Wang, Wu-Hsiung Wu, Stuart Brown, D Frank Hsu, Christina Schweikert, Zuojian Tang, Chien-Yu Chen, Jer-Wei Chang, et al. Applications in Cancer-Related Research. *Systems Biology*, 10, 9789814324465_0001.

Karolchik, Donna, Angie S Hinrichs, and W James Kent (2009). The UCSC genome browser. *Current protocols in bioinformatics*, pp. 1–4.

Kc, Dukka B and Dennis R Livesay (2011). Topology improves phylogenetic motif functional site predictions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8 (1), pp. 226–233.

Keilwagen, Jens, Jan Grau, Ivan A Paponov, Stefan Posch, Marc Strickert, and Ivo Grosse (2011). De-novo discovery of differentially abundant transcription factor binding sites including their positional preference. *PLoS Comput Biol*, 7 (2), e1001070.

Kel, Alexander E, Ellen Gößling, Ingmar Reuter, Evgeny Cheremushkin, Olga V Kel-Margoulis, and Edgar Wingender (2003). MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic acids research*, 31 (13), pp. 3576–3579.

Kulakovskiy, I. V., V. A. Boeva, A. V. Favorov, and V. J. Makeev (2010). Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, 26 (20), pp. 2622–2623.

Kulakovskiy, Ivan, Victor Levitsky, Dmitry Oshchepkov, Leonid Bryzgalov, Ilya Vorontsov, and Vsevolod Makeev (2013). From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *Journal of bioinformatics and computational biology*, 11 (01), p. 1340004.

Lawrence, Charles E, Stephen F Altschul, Mark S Boguski, Jun S Liu, Andrew F Neuwald, and John C Wootton (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *science*, 262 (5131), pp. 208–214.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey (2013). Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*, 9 (8), e1003118+.

Lee, Hsin-Tsang, Derek Leonard, Xiaoming Wang, and Dmitri Loguinov (2009). IRLbot: scaling to 6 billion pages and beyond. *ACM Transactions on the Web (TWEB)*, 3 (3), p. 8.

Luco, Reini F, Qun Pan, Kaoru Tominaga, Benjamin J Blencowe, Olivia M Pereira-Smith, and Tom Misteli (2010). Regulation of alternative splicing by histone modifications. *Science*, 327 (5968), pp. 996–1000.

Ma, Wenxiu, William S Noble, and Timothy L Bailey (2014). Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nature protocols*, 9 (6), pp. 1428–1450.

Maddelein, Davy, Niklaas Colaert, Iain Buchanan, Niels Hulstaert, Kris Gevaert, and Lennart Martens (2015). The iceLogo web server and SOAP service for determining protein consensus sequences. *Nucleic acids research*, gkv385.

Mahony, Shaun and Panayiotis V Benos (2007). STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic acids research*, 35 (suppl 2), W253–W258.

Malone, John and Brian Oliver (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9 (1), pp. 34+.

Mar-Aguilar, Fermín, Cristina Rodríguez-Padilla, and Diana Reséndez-Pérez (2016). Web-based tools for microRNAs involved in human cancer (Review). *Oncology Letters*, 11 (6), pp. 3563–3570.

Matys, V., E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender (2003). TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31 (1), pp. 374–378.

Maza, Michael de la and Bruce Tidor (1993). "An Analysis of Selection Procedures with Particular Attention Paid to Proportional and Boltzmann Selection". In: *Proceedings of the 5th International Conference on Genetic Algorithms*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 124–131.

Miller, Neil A, Emily G Farrow, Margaret Gibson, Laurel K Willig, Greyson Twist, Byunggil Yoo, Tyler Marrs, Shane Corder, Lisa Krivohlavek, Adam Walter, et al. (2015). A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome medicine*, 7 (1), p. 1.

Moses, Alan, Derek Chiang, Daniel Pollard, Venky Iyer, and Michael Eisen (2004). MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biology*, 5 (12), R98.

Nelson, Peter T, Wang-Xia Wang, and Bernard W Rajeev (2008). MicroRNAs (miRNAs) in neurodegenerative diseases. *Brain Pathology*, 18 (1), pp. 130–138.

Neph, Shane and Martin Tompa (2006). MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes. *Nucleic acids research*, 34 (suppl 2), W366–W368.

Nettling, Martin, Nils Thieme, Andreas Both, and Ivo Grosse (2014). DRUMS: Disk Repository with Update Management and Select option for high throughput sequencing data. *BMC bioinformatics*, 15 (1), p. 1.

Nettling, Martin, Hendrik Treutler, Jesus Cerquides, and Ivo Grosse (2016). Detecting and correcting the binding-affinity bias in ChIP-seq data using inter-species information. *BMC genomics*, 17 (1), p. 1.

— (2017a). Combining phylogenetic footprinting with motif models incorporating intra-motif dependencies. *BMC bioinformatics*, 18 (1), p. 141.

— (2017b). Unrealistic phylogenetic trees may improve phylogenetic footprinting. *Bioinformatics*, p. 8.

Nettling, Martin, Hendrik Treutler, Jan Grau, Jens Keilwagen, Stefan Posch, and Ivo Grosse (2015). DiffLogo: a comparative visualization of sequence motifs. *BMC bioinformatics*, 16 (1), p. 1.

Neubeck, Lis, Nicole Lowres, Emelia J Benjamin, S Ben Freedman, Genevieve Coorey, and Julie Redfern (2015). The mobile revolution [mdash] using smartphone apps to prevent cardiovascular disease. *Nature Reviews Cardiology*, 12 (6), pp. 350–360.

Newberg, Lee A, William A Thompson, Sean Conlan, Thomas M Smith, Lee Ann McCue, and Charles E Lawrence (2007). A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics*, 23 (14), pp. 1718–1727.

Nguyen, Tung and Ioannis Androulakis (2009). Recent Advances in the Computational Discovery of Transcription Factor Binding Sites. *Algorithms*, 2 (1), pp. 582–605.

Noble, William Stafford (2009). A quick guide to organizing computational biology projects. *PLoS Comput Biol*, 5 (7), e1000424.

O'Grady, Stephen (2015). *The RedMonk Programming Language Rankings: January 2015.* URL: `http://redmonk.com/sogrady/2015/01/14/language-rankings-1-15/` (visited on 12/10/2016).

Ozsolak, Fatih, Laura L Poling, Zhengxin Wang, Hui Liu, X Shirley Liu, Robert G Roeder, Xinmin Zhang, Jun S Song, and David E Fisher (2008). Chromatin structure analyses identify miRNA promoters. *Genes & development*, 22 (22), pp. 3172–3183.

Pang, Linsey Xiaolin, Sanjay Chawla, Wei Liu, and Yu Zheng (2013). On detection of emerging anomalous traffic patterns using GPS data. *Data & Knowledge Engineering*, 87, pp. 357–373.

Park, Daechan, Yaelim Lee, Gurvani Bhupindersingh, and Vishwanath R Iyer (2013). Widespread Misinterpretable ChIP-seq Bias in Yeast. *PloS one*, 8 (12), e83506.

Redhead, Emma and Timothy L Bailey (2007). Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC bioinformatics*, 8 (1), p. 1.

Reik, Wolf (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447 (7143), pp. 425–432.

Rigden, Daniel J, Xosé M Fernández-Suárez, and Michael Y Galperin (2016). The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection. *Nucleic acids research*, 44 (D1), pp. D1–D6.

Rigden, Daniel J., Xosé M. Fernández-Suárez, and Michael Y. Galperin (2016). The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection. *Nucleic Acids Research*, 44 (D1), pp. D1–D6.

Ross, Michael G., Carsten Russ, Maura Costello, Andrew Hollinger, Niall J. Lennon, Ryan Hegarty, Chad Nusbaum, and David B. Jaffe (2013). Characterizing and measuring bias in sequence data. *Genome Biol*, 14 (5), R51.

Salama, Rafik A and Dov J Stekel (2010). Inclusion of neighboring base interdependencies substantially improves genome-wide prokaryotic transcription factor binding site prediction. *Nucleic acids research*, 38 (12), e135–e135.

Schneider, T. D. and R. M. Stephens (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18 (20), pp. 6097–6100.

Siddharthan, Rahul (2008). PhyloGibbs-MP: module prediction and discriminative motif-finding by Gibbs sampling. *PLoS Comput Biol*, 4 (8), e1000156.

Siddharthan, Rahul, Eric D. Siggia, and Erik van Nimwegen (2005). PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny. *PLoS Comput Biol*, 1 (7), e67+.

Siebert, Matthias and Johannes Söding (2016). Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*, 44 (13), pp. 6055–6069.

Single Nucleotide Polymorphism (2012).

Sinha, Saurabh, Mathieu Blanchette, and Martin Tompa (2004). PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC bioinformatics*, 5 (1), p. 170.

Slotkin, R Keith and Robert Martienssen (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 8 (4), pp. 272–285.

Small, Eric M and Eric N Olson (2011). Pervasive roles of microRNAs in cardiovascular biology. *Nature*, 469 (7330), pp. 336–342.

Sokal, Robert R (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, 38, pp. 1409–1438.

Sudmant, Peter H, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, et al. (2015).

An integrated map of structural variation in 2,504 human genomes. *Nature*, 526 (7571), pp. 75–81.

Sultan, Marc, Marcel H Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, Tatjana Borodina, Aleksey Soldatov, Dmitri Parkhomchuk, et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321 (5891), pp. 956–960.

Tam, Oliver H, Alexei A Aravin, Paula Stein, Angelique Girard, Elizabeth P Murchison, Sihem Cheloufi, Emily Hodges, Martin Anger, Ravi Sachidanandam, Richard M Schultz, et al. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453 (7194), pp. 534–538.

Teytelman, Leonid, Deborah M. Thurtle, Jasper Rine, and Alexander van Oudenaarden (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences*, 110 (46), pp. 18602–18607.

Tippmann, Sylvia et al. (2015). Programming tools: Adventures with R. *Nature*, 517 (7532), pp. 109–110.

Tompa, Martin, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W James Kent, et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, 23 (1), pp. 137–144.

Tran, Ngoc Tam L and Chun-Hsi Huang (2014). A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biol Direct*, 9 (1), p. 4.

Vacic, Vladimir, Lilia M Iakoucheva, and Predrag Radivojac (2006). Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, 22 (12), pp. 1536–1537.

Wang, Jie, Jiali Zhuang, Sowmya Iyer, XinYing Lin, Troy W Whitfield, Melissa C Greven, Brian G Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research*, 22 (9), pp. 1798–1812.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10 (1), pp. 57–63.

Watson, James D, Francis HC Crick, et al. (1953). Molecular structure of nucleic acids. *Nature*, 171 (4356), pp. 737–738.

Wheeler, Travis J and John D Kececioglu (2007). Multiple alignment by aligning alignments. *Bioinformatics*, 23 (13), pp. i559–i568.

Wilbanks, Elizabeth G and Marc T Facciotti (2010). Evaluation of algorithm performance in ChIP-seq peak detection. *PloS one*, 5 (7), e11471.

Wilson, Greg, DA Aruliah, C Titus Brown, Neil P Chue Hong, Matt Davis, Richard T Guy, Steven HD Haddock, Katy Huff, Ian M Mitchell, Mark D Plumbley, et al. (2014). Best practices for scientific computing. *PLoS biology*, 12 (1), e1001745.

Zambelli, Federico, Graziano Pesole, and Giulio Pavesi (2012). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in bioinformatics*, bbs016.

4 Data acquisition and data preparation

4.1 miRGen 2.0: a database of microRNA genomic information and regulation

P Alexiou, T Vergoulis, **M Gleditzsch**, G Prekas, T Dalamagas, M Megraw, I Grosse, T Sellis, AG Hatzigeorgiou. 2009. miRGen 2.0: a database of microRNA genomic information and regulation. *Nucl. Acids Res.* 38 (suppl 1): D137-D141. *doi:10.1093/nar/gkp888*

Published online 22 October 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D137–D141***
doi:10.1093/nar/gkp888

miRGen 2.0: a database of microRNA genomic information and regulation

Panagiotis Alexiou[1,2,*], Thanasis Vergoulis[3,4], Martin Gleditzsch[5], George Prekas[4], Theodore Dalamagas[3], Molly Megraw[6], Ivo Grosse[5], Timos Sellis[3,4] and Artemis G. Hatzigeorgiou[1,7,*]

[1]Institute of Molecular Oncology, Biomedical Sciences Research Center 'Alexander Fleming', Vari, [2]School of Biology, Aristotle University of Thessaloniki, Thessaloniki, [3]Institute for the Management of Information Systems, ''Athena'' Research Center, [4]Knowledge and Database Systems Lab, Department of Computer Science, School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece, [5]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle, Germany, [6]Institute for Genome Sciences and Policy, Duke University, Durham, NC and [7]Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA, USA

Received September 15, 2009; Accepted October 4, 2009

ABSTRACT

MicroRNAs are small, non-protein coding RNA molecules known to regulate the expression of genes by binding to the 3′UTR region of mRNAs. MicroRNAs are produced from longer transcripts which can code for more than one mature miRNAs. miRGen 2.0 is a database that aims to provide comprehensive information about the position of human and mouse microRNA coding transcripts and their regulation by transcription factors, including a unique compilation of both predicted and experimentally supported data. Expression profiles of microRNAs in several tissues and cell lines, single nucleotide polymorphism locations, microRNA target prediction on protein coding genes and mapping of miRNA targets of co-regulated miRNAs on biological pathways are also integrated into the database and user interface. The miRGen database will be continuously maintained and freely available at http://www.microrna.gr/mirgen/.

INTRODUCTION

MicroRNAs (miRNAs) are single-stranded non-coding RNA molecules of ~21 nucleotides in length, that function as regulators of gene expression by binding to messenger RNA (mRNA) molecules and destabilizing them or inhibiting their translation. They are found to be implicated in a wide range of physiological molecular processes, and their deregulation leads to diverse diseases (1–3).

MiRNAs are located in intergenic regions or in the introns of protein coding genes. They are transcribed by RNA Polymerase II as independent transcripts or as part of the transcript of a host gene. Only a small group of miRNAs located inside ALU repetitive elements is transcribed by RNA Polymerase III. A miRNA transcript can host more than one miRNA and can be several thousand nucleotides long including introns.

A promoter region is located around the transcription start site (TSS) of a transcript and is regulated by proteins that bind to this region. Evidence thus far suggests that binding sites for transcription factors (TFs) are similarly distributed within the promoters of both protein coding genes and miRNA transcripts (4). MiRNA primary transcripts (pri-miRNA) are processed in the nucleus to form pre-miRNAs, ~70-nucleotide stem–loop structures also called miRNA hairpins. These are later processed into mature miRNAs in the cytoplasm via interaction with the endonuclease Dicer, which also initiates the formation of the RNA-induced silencing complex (RISC). Since primary transcripts are short lived and present only inside the nucleus, it is hard to identify them with standard molecular techniques.

After the Dicer enzyme cleaves the pre-miRNA stem–loop, two complementary short RNA molecules are formed, but only one of them—the guiding strand—is predominantly integrated into the RISC complex.

*To whom correspondence should be addressed. Tel: +30 210 9656310 (int. 248); Email: pan.alexiou@fleming.gr
Correspondence may also be addressed to Artemis G. Hatzigeorgiou. Tel: +30 210 9656310 (int. 190); Fax: +30 210 9653934; Email: hatzigeorgiou@fleming.gr

D138 Nucleic Acids Research, 2010, Vol. 38, Database issue

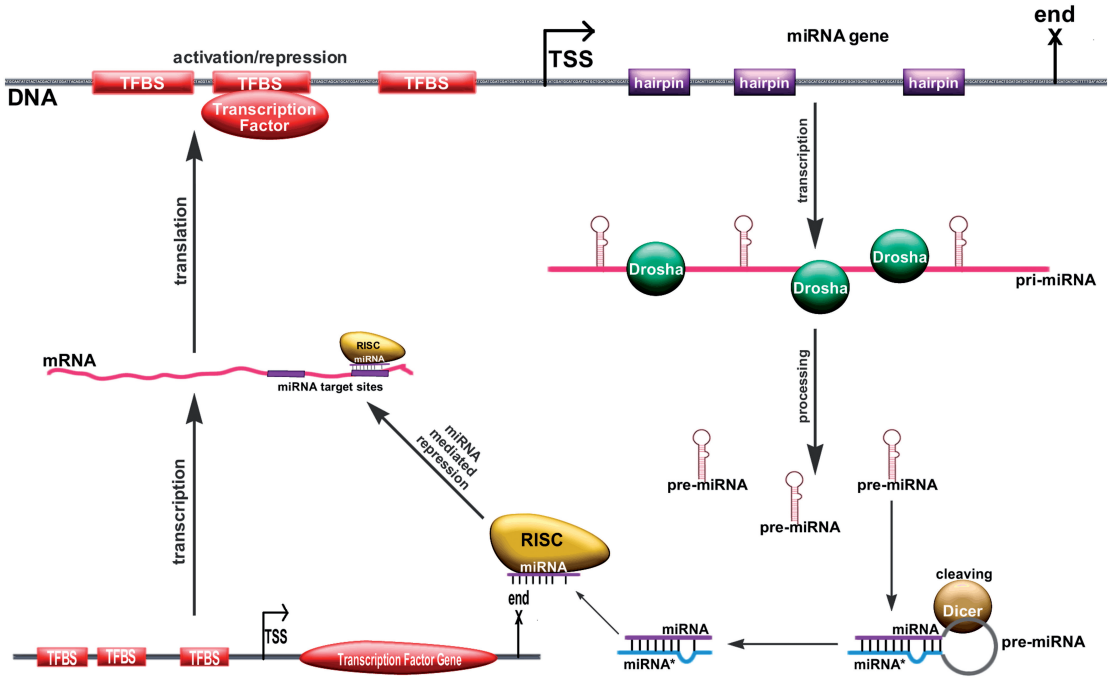


Figure 1. A miRNA gene (top) is controlled by several TFs whose binding sites (TFBSs) are located near the TSS of this gene. When transcribed, the miRNA gene produces a long pri-miRNA molecule. The pri-miRNA molecule is cleaved by Drosha and yields the pre-miRNA stem-loop (hairpin) structure. The enzyme Dicer cleaves the loop part of the hairpin and produces the miRNA-miRNA* duplex. One chain of the miRNA duplex is incorporated into the RISC complex and can regulate mRNA translation by binding in a sequence specific manner to the 3′UTR region of mRNAs. In this example, the miRNA (produced after a TF binds to its promoter) regulates the translation of the promoter in a typical negative feedback control loop.

The remaining strand, known as the miRNA*, anti-guide or passenger strand, is usally degraded. However, the proportion of the integration of each strand varies with the miRNA species, with some miRNAs having almost equal abundance of each of the two strands incorporated into RISC. Another common nomenclature for complementary miRNA strands is the –3p and –5p naming convention—these names do not imply which miRNA is more commonly incorporated to the RISC complex. The miRNA–miRNA* and miRNA-3p–miRNA-5p nomenclatures are both widely used in the community, often to denote the same complementary miRNA pair. Mature miRNA molecules are bound by the RISC complex, are guided to specific motifs within the 3′UTR of protein coding mRNAs, and prevent these mRNAs from being translated into protein. The biogenesis of miRNAs and their regulation by TFs is diagrammed in Figure 1.

Single-nucleotide polymorphisms (SNPs) are DNA sequence positions at which a single nucleotide varies between individuals of the same species. SNPs are fairly common in mammalian genomes (the human genome contains ~20 million SNP sites) and have been extensively linked to genetic abnormalities and disease (5).

In the previous version of the miRGen database (6), co-expressed miRNA clusters were identified based on their distance and genomic features surrounding them. With the availability of experimental data we were able, in miRGen 2.0, to mine prominent literature sources that identify miRNA primary transcripts in mammals (human and mouse genomes). Moreover, we have mapped TF binding sites (TFBSs) within the regions upstream of these miRNA primary transcript TSSs and incorporated expression profiles of miRNAs in several tissues, the mapping of SNPs within genomic locations of miRNA hairpins and the mapping of SNPs within the TFBSs found upstream of miRNA genes. The interplay of these different information sources concerning genomic features associated with miRNA genes and their expression levels can be used to study the function of miRNAs and their deregulation in disease. For instance, a user interested in a specific TF can find miRNA genes associated with this TF, find the expression levels of these miRNAs in a possible tissue of interest, possibly find some SNPs on the TFBSs or the miRNA locations on the genome that relate to a possible disease of interest and finally find predicted targets of the miRNAs associated with the TF of interest, and molecular pathways in which the targets of each of these miRNAs separately or together are implicated.

4. DATA ACQUISITION AND DATA PREPARATION

DATA GENERATION

miRNA coding transcripts

MiRNA transcripts in human and mouse were identified from four literature sources:

(i) Corcoran *et al.* (7) used PolII immunoprecipitation data and ChIP–chip on lung epithelial cells to identify miRNA transcripts and their promoter regions.
(ii) Landgraf *et al.* (8) sequenced 250 small RNA libraries corresponding to 26 different organ systems and cell types of human and mouse, with ~1000 miRNA clones per library and identified miRNA coding genes. In this study the whole transcripts of miRNA coding genes were identified, as well as protein coding genes that contain miRNAs.
(iii) Oszolak *et al.* (9) predicted the location of the proximal promoters of human miRNAs by combining nucleosome mapping with promoter chromatin signatures in MALME, HeLa and UACC62 cells. Although the TSS of miRNA genes was identified in this study, the end of the transcript was not provided. We have provided end of the last miRNA that is a member of a gene as an approximation of the transcript end.
(iv) Marson *et al.* (10) used ChIP-seq data to identify promoters of miRNA genes in embryonic stem cells. They identified promoters and co-regulated miRNAs, but the exact position of the TSS was not identified. For this reason we have used the start of the first miRNA of each cluster as the putative TSS. Additionally, coordinates provided by Marson *et al.* had to be lifted over using 'UCSC lift over tool' to the current genome build (hg18, mm9). In cases where putative rather than experimentally verified positions are used, they are denoted in the graphical interface as 'computational TSS'.

In total, 812 human miRNA coding transcripts and 386 mouse miRNA coding transcripts were identified. Of them, 423 were shown in the corresponding papers to be associated with protein coding genes (intragenic miRNA transcripts). More than one of the above publications have usually identified transcripts corresponding to a miRNA. When this is the case, transcripts from all methods are returned to the user.

Since these studies were published, additional miRNAs have been identified. When novel miRNAs are located within the coordinates of clusters given by any of these publications, this miRNA is added to the cluster. For names that changed or were given differently than the current standard, manual curation with reference to mirBase (11) was used to identify and replace these names according to the current standard. For all the above reasons it is possible that the number of genes used in miRGen (Table 1) does not correspond perfectly to the number stated in the corresponding publications.

Table 1. Number of miRNA coding genes and mature miRNAs identified in each of the experimental studies used to populate the miRGen database

References	Human Genes	Human miRNA	Mouse Genes	Mouse miRNA
Corcoran *et al.* (7)	73	148	–	–
Landgraf *et al.* (8)	201	347	191	590
Ozsolak *et al.* (9)	191	268	–	–
Marson *et al.* (10)	346	507	195	422

TFBS identification

In order to determine putative TFBSs near the TSS of miRNA primary transcripts, we used the freely available tool MatchTM (12). MatchTM uses the public library of position weight matrices from Transfac 6.0—cite: TRANSFAC: an integrated system for gene expression regulation. We matched all vertebrate TF matrices to the regions spanning from 5 kb upstream of each TSS to 1 kb downstream of the TSS. As criterion for determining the cut-off values we chose the minimization of false positives in order to produce a strict set of predictions without too many falsely predicted TFBSs. Two scores are calculated for each putative TFBS. The matrix similarity score describes the quality of a match between a whole matrix and an arbitrary part of the input sequences. Analogously, the core similarity score denotes the quality of the match between the core sequence of a matrix (i.e. the five most conserved positions within a matrix) and a part of the input sequence.

miRNA expression profiles

miRNA expression profiles were identified from the mammalian miRNA expression atlas (8). Information for the expression profiles of 548 human and 451 mouse miRNAs over 172 human and 68 mouse small RNA libraries were derived from cell lines and tissues.

SNPs

SNPs located within the genomic positions of miRNA hairpins and corresponding TFBSs were downloaded from the UCSC table browser (13). For human, Polymorphism data from dbSnp database (14) or genotyping arrays SNP130 were used with 18 833 531 identified SNPs. For mouse, SNP128 was used with 14 893 502 identified SNPs.

Implementations

The miRGen repository has been implemented using relational database technology. All data are stored in a MySQL relational database management system. Figure 2 illustrates part of the entity-relationship model of our application. All results are available through a user-friendly interface that allows searches for miRNAs and for TFs of interest. For mature miRNAs, it is possible to view targets predicted by the program microT-ANN and for miRNAs found in the same transcript, the user can see a functional annotation of their targets on molecular

D140 *Nucleic Acids Research, 2010, Vol. 38, Database issue*

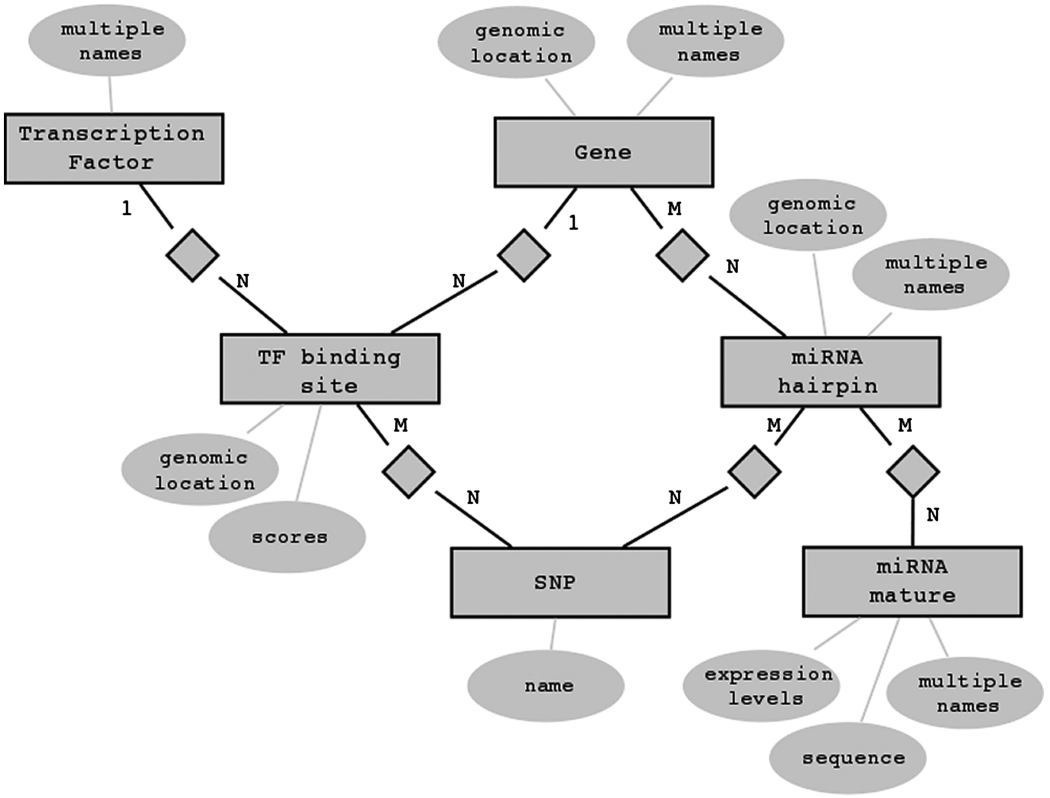


Figure 2. The miRGen database schema. TFs (top right) bind through TF binding sites to miRNA genes. miRNA genes (top) contain miRNA hairpins that signify the genomic location of the mature miRNA-miRNA* duplex. miRNA hairpins are processed into mature miRNAs. Usually, one miRNA hairpin produces two mature miRNAs, but a mature miRNA can be produced by more than one hairpin in different genomic locations. Both TFBSs and miRNA hairpins are genomic features that can contain SNPs. Mature miRNAs are associated with their expression levels in different tissues and cell types.

pathways through the application DIANA-mirPath (15). Figure 3 shows an overview of the interface and highlights links to external databases—UCSC genome browser (13), iHop (16), dbSNP (14), mirBase (11).

DISCUSSION

This version of miRGen is the first attempt to build a widely accessible and user-friendly database that connects TFs and miRNAs through putative and experimentally supported functional relationships. The connections identified in the database will further our understanding of the TF-mediated regulation of miRNA genes, and pave the way for the mapping of the interplay between TFs and miRNAs as regulatory molecules. The identification of SNPs on miRNA locations and their corresponding TFBSs, as well as the expression profiles of miRNAs can improve our insight into the involvement of miRNAs in developmental processes and disease.

Deregulation of TF-mediated gene expression has been shown to extensively affect protein coding genes, and lead to disease (17,18). MiRNA expression levels have also been shown to change significantly in different disease states (19,20). The availability of both these resources in the same database will allow researchers to identify regulatory elements, such as TFs that may affect the expression of miRNAs. For this reason, we believe miRGen 2.0 will be an important resource for researchers of diverse disciplines interested in miRNA regulation and function.

AVAILABILITY

The miRGen database will be continuously maintained and freely available at http://www.microrna.gr/mirgen/.

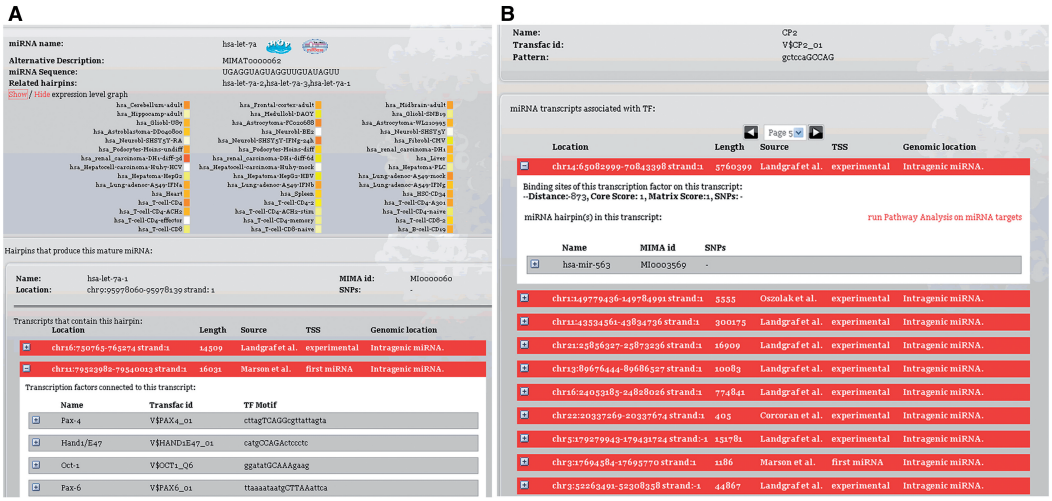


Figure 3. The user is able to query the database either by miRNA name, or by the name of the TF of interest. When a miRNA search is performed (Figure 3a), all distinct locations on the genome (hairpins) that could code for this miRNA are returned, and the user can see details for any of the possible overlapping transcripts identified for each location, usually predicted by different papers. Each transcript tab contains information about TFBSs located from 5 kb upstream to 1 kb downstream of the transcript start. Additionally, information on the expression levels of the mature miRNA are displayed as a heat map. Searching for a TF of interest (Figure 3b) returns all miRNA coding genes for which at least one binding site for this TF is found. Information on the gene, the TFBSs, and the mature miRNAs coded for by the gene can be seen in tabs. All instances of TFBSs and miRNA hairpins are associated with corresponding SNPs mapping on their genomic locations. For all transcripts, the literature source of the gene is displayed, the identification of the TSS (experimental if the TSS was identified in the paper, computational if it was calculated by computational means and first miRNA if the start of the first miRNA serves as a substitute for an unknown TSS), and whether the gene is intragenic or is co-expressed with a protein-coding gene.

FUNDING

Aristeia Award from General Secretary Research and Technology, Greece. Funding for open access charge: The Aristeia Award from General Secretary Research and Technology, Greece.

Conflict of interest statement. None declared.

REFERENCES

1. Gartel,A.L. and Kandel,E.S. (2008) miRNAs: little known mediators of oncogenesis. *Semin. Cancer Biol.*, **18**, 103–110.
2. Fabbri,M., Croce,C.M. and Calin,G.A. (2009) MicroRNAs in the ontogeny of leukemias and lymphomas. *Leuk Lymphoma*, **50**, 160–170.
3. Latronico,M.V., Catalucci,D. and Condorelli,G. (2008) MicroRNA and cardiac pathologies. *Physiol. Genomics*, **34**, 239–242.
4. Megraw,M., Baev,V., Rusinov,V., Jensen,S.T., Kalantidis,K. and Hatzigeorgiou,A.G. (2006) MicroRNA promoter element discovery in Arabidopsis. *RNA*, **12**, 1612–1619.
5. Brookes,A.J. (1999) The essence of SNPs. *Gene*, **234**, 177–186.
6. Megraw,M., Sethupathy,P., Corda,B. and Hatzigeorgiou,A.G. (2007) miRGen: a database for the study of animal microRNA genomic organization and function. *Nucleic Acids Res.*, **35**, D149–D155.
7. Corcoran,D.L., Pandit,K.V., Gordon,B., Bhattacharjee,A., Kaminski,N. and Benos,P.V. (2009) Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. *PLoS ONE*, **4**, e5279.
8. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
9. Ozsolak,F., Poling,L.L., Wang,Z., Liu,H., Liu,X.S., Roeder,R.G., Zhang,X., Song,J.S. and Fisher,D.E. (2008) Chromatin structure analyses identify miRNA promoters. *Genes Dev.*, **22**, 3172–3183.
10. Marson,A., Levine,S.S., Cole,M.F., Frampton,G.M., Brambrink,T., Johnstone,S., Guenther,M.G., Johnston,W.K., Wernig,M., Newman,J. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.
11. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
12. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
13. Karolchik,D., Hinrichs,A.S. and Kent,W.J. (2007) The UCSC Genome Browser. *Curr. Protoc. Bioinformatics*, Chapter 1, Unit 14.
14. Smigielski,E.M., Sirotkin,K., Ward,M. and Sherry,S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.*, **28**, 352–355.
15. Papadopoulos,G.L., Alexiou,P., Maragkakis,M., Reczko,M. and Hatzigeorgiou,A.G. (2009) DIANA-mirPath: integrating human and mouse microRNAs in pathways. *Bioinformatics*, **25**, 1991–1993.
16. Fernandez,J.M., Hoffmann,R. and Valencia,A. (2007) iHOP web services. *Nucleic Acids Res.*, **35**, W21–W26.
17. Karin,M. (2006) Nuclear factor-kappaB in cancer development and progression. *Nature*, **441**, 431–436.
18. Maiese,K., Chong,Z.Z., Shang,Y.C. and Hou,J. (2008) Clever cancer strategies with FoxO transcription factors. *Cell Cycle*, **7**, 3829–3839.
19. Nikiforova,M.N., Chiosea,S.I. and Nikiforov,Y.E. (2009) MicroRNA expression profiles in thyroid tumors. *Endocr. Pathol.*, **20**, 85–91.
20. Aslam,M.I., Taylor,K., Pringle,J.H. and Jameson,J.S. (2009) MicroRNAs are novel biomarkers of colorectal cancer. *Br. J. Surg.*, **96**, 702–710.

4.2 DRUMS: Disk Repository with Update Management and Select option for high throughput sequencing data

M Nettling, N Thieme, A Both, I Grosse. 2014. DRUMS: Disk Repository with Update Management and Select option for high throughput sequencing data. *BMC bioinformatics*, 15:1. *doi:10.1186/1471-2105-15-38*

Nettling *et al. BMC Bioinformatics* 2014, **15**:38
http://www.biomedcentral.com/1471-2105/15/38

BMC Bioinformatics

SOFTWARE Open Access

DRUMS: Disk Repository with Update Management and Select option for high throughput sequencing data

Martin Nettling[1,2], Nils Thieme[2], Andreas Both[2*] and Ivo Grosse[1,3]

Abstract

Background: New technologies for analyzing biological samples, like next generation sequencing, are producing a growing amount of data together with quality scores. Moreover, software tools (e.g., for mapping sequence reads), calculating transcription factor binding probabilities, estimating epigenetic modification enriched regions or determining single nucleotide polymorphism increase this amount of position-specific DNA-related data even further. Hence, requesting data becomes challenging and expensive and is often implemented using specialised hardware. In addition, picking specific data as fast as possible becomes increasingly important in many fields of science. The general problem of handling big data sets was addressed by developing specialized databases like HBase, HyperTable or Cassandra. However, these database solutions require also specialized or distributed hardware leading to expensive investments. To the best of our knowledge, there is no database capable of (i) storing billions of position-specific DNA-related records, (ii) performing fast and resource saving requests, and (iii) running on a single standard computer hardware.

Results: Here, we present DRUMS (Disk Repository with Update Management and Select option), satisfying demands (i)-(iii). It tackles the weaknesses of traditional databases while handling position-specific DNA-related data in an efficient manner. DRUMS is capable of storing up to billions of records. Moreover, it focuses on optimizing relating single lookups as range request, which are needed permanently for computations in bioinformatics. To validate the power of DRUMS, we compare it to the widely used MySQL database. The test setting considers two biological data sets. We use standard desktop hardware as test environment.

Conclusions: DRUMS outperforms MySQL in writing and reading records by a factor of two up to a factor of 10000. Furthermore, it can work with significantly larger data sets. Our work focuses on mid-sized data sets up to several billion records without requiring cluster technology. Storing position-specific data is a general problem and the concept we present here is a generalized approach. Hence, it can be easily applied to other fields of bioinformatics.

Keywords: Database, HERV, SNP, DNA related data, High throughput data

Background

With the beginning of the information age in the 90s of the last century, a large set of processes are established to manipulate and analyze data. In particular in the field of bioinformatics, many different workflows produce a growing amount of data. One example are sequencing technologies, which are capable of sequencing an entire human genome in less than a day. Moreover, extensive software suites for analyzing biological data sets exist, e.g. http://galaxy.psu.edu/ [1-3]. In addition, it is possible that an analyzing process produces more output data than provided input. For example, the input size of the HERV data set used in this work is about 4 GB. The output of the mapping with BLAST is about 50 GB large. Hence, rapid processes for storing and querying data are needed as it has impact on the general performance of the analytic processes.

*Correspondence: andreas.both@unister.de
[2]R&D, Unister GmbH, Leipzig, Germany
Full list of author information is available at the end of the article

Position-specific DNA related data (psDrd)

In the field of bioinformatics, data related to DNA sequences are of particular importance. Examples are single nucleotide polymorphisms (SNPs) [4], transcription factor binding affinities and probabilities [5,6], and RNAseq data [7,8]. We generalize these types of data by the term position-specific DNA-related data (*psDrd*). A *psDrd* record is an information related to a specific DNA position. *psDrd* records have three characteristics. First, a *psDrd* record R can be represented by a key-value pair $R = (K, V)$. The key K is composed of the sequence identifier and the position of the associated value V. Hence, the key is unique, and records can be easily sorted. Second, *psDrd* records are usually requested by region (e.g., querying for all mutations in a specific gene or looking for transcription factors that are binding near a given position). We call this kind of access *range select*. Third, all *psDrd* of the same kind need the same space to be stored on device, i.e., two different records are represented by the same number of bytes. In contrast, textual annotations are generally of variable length. These three specific properties can be utilized for optimizing data handling of *psDrd*.

Time- and resource-intensive computations on psDrd

Many biological processes and bioinformatics algorithm have *psDrd* as input or output. This type of data is essential for understanding biological and biochemical processes. Furthermore, diagnostics in medicine for cancer prediction and genetic diseases are using *psDrd* [9-11].

Many activities in bioinformatics focus on analyzing *psDrd*. However, often file and folder strategy or a standard databases like MySQL [12] are used for data management. These approaches are straightforward but not optimized for the intended processing of *psDrd*. In addition, data types used in these tools are expensive and might lead to an exhaustive usage of valuable resources [13]. Both problems lead to resource-intensive requests of *psDrd*. For example, when performing range selects using MySQL, nearly each record in the range must be fetched by a costly random access to the storage. Because of the limits of standard desktop hardware, this might cause a bottleneck during data processing.

Requirements

The following requirements result from the above mentioned problems: The data management must be usable with standard desktop technology. It must be possible to store billions of data records. Platform independency was defined as an additional requirement (derived from the well-known segmentation of operation systems). Handling massive read requests during analytic processes has to be possible. While optimizing data handling of *psDrd*, the three specific properties from section *"Position-specific DNA related data (psDrd)"* have to be obeyed.

Implementation

In this section, we first describe a concept called DRUM, on which DRUMS is based. Subsequently, we describe the architecture of DRUMS. Finally, we briefly sketch the implementation of DRUMS in Java considering the three main requirements of handling *psDrd* data sets efficiently.

DRUM concept

The DRUM (Disk Repository with Update Management) concept [14] allows to store large collections of key-value pairs (KVs). DRUM allows fast bulk inserts without generating duplicate entries. To enable fast processing, incoming *psDrd* records (K, V) are allocated based on their key K to separate buffers B in the main memory: $\mathcal{M}(K) \rightarrow B_i$. Those buffers are continuously written to their counterparts on disk (D), where they are called *buckets*. If a bucket on disk reaches a predefined size, a synchronisation process with the persistently saved data (on the hard disk) starts. The process is executed in the following way: A disk bucket is entirely read to a disk cache. There it is sorted. Thereafter, a synchronisation is performed by combining each bucket after the other with the corresponding cache. As the records of the disk cache are also sorted, using mergesort is efficient. The synchronisation process is blocking all other processes within DRUM.

The DRUM concept is very suitable for storing *psDrd*. However, requesting data efficiently was never a goal of this approach. Hence, neither single lookups nor range selects have been optimized. Furthermore, when synchronisation is performed, DRUM is not able to receive and cache new *psDrd* records. In the following, we propose an extension of DRUM that addresses these shortcomings.

Extensions by the DRUMS concept

We extend the DRUM concept by allowing the selection of records by key (*single lookup*) or by range (*range selects*). Within this concept we decoupled I/O-processes from memory processes to avoid blocking single components.

Following the three *psDrd* data properties, the following architecture decisions were made for DRUMS in addition to the DRUM concept: 1) All records are equally sized, so that jumping to the start position of an arbitrary record in the file is possible. Therefore, a sparse index [15] can be applied efficiently, making rapid single selects possible by the following two steps: The sparse index points to a block of records, where the *psDrd* of interest might be found. To finally find the requested record, a binary search is performed. The binary search massively benefits from equally sized records. 2) Records, which are close to each other on DNA are stored close on disk according to their keys. This enables efficient range selects. 3) Records are organized in

Nettling *et al. BMC Bioinformatics* 2014, **15**:38
http://www.biomedcentral.com/1471-2105/15/38

buckets and chunks, which permits efficient prefiltering of regions of interest within a bucket.

Architecture of DRUMS

DRUMS is composed of the interacting components described in this section. Before each component is described in detail, we give a high-level overview of the insert and select process of DRUMS.

Processes

Insert process The high-level overview of the insert process of DRUMS is shown in Figure 1. KV pairs are sent to DRUMS. As in DRUM, the incoming records are already distributed in memory between n buffers B (called memory buckets). Each bucket B_i in memory has a corresponding bucket D_i on disk. The sizes of the buckets are dynamic. If a bucket B_i exceeds a predefined size or memory limitations are reached, a synchronisation process, consisting of four phases, is started:

1) The bucket B_i is taken and replaced by an empty one. Hence, incoming data can still be buffered. 2) The KV pairs of B_i are sorted by their keys. 3) B_i and D_i are synchronised using mergesort. Already existing records can be updated using state-dependent operations. 4) The merged data is continuously written back to bucket D_i. Hence, input data is now saved persistently on the disk.

Note: Step 3 and 4 of the synchronization process are performed chunk-wise, so that optimal read and write performance can be achieved. The optimal chunk-size depends on the used hardware, the size of a single record, the expected data volume, and several parameters in DRUMS. Therefore, it has to be determined empirically.

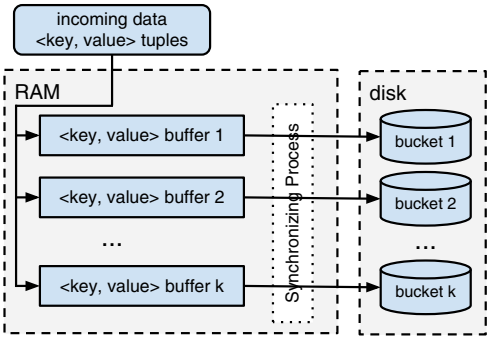


Figure 1 High level overview of insert process. Key-value pairs are sent to DRUMS. The incoming records are distributed between k buffers (memory buckets), based on their key. If a bucket B_i exceeds a predefined size or memory limitations are reached, a synchronisation process is instantiated.

Range select process Figure 2 shows the high-level overview of the select process. When a request is sent to DRUMS, four steps are performed to read the requested records given by the keys K_S and K_E (start and end of the range). 1) The requested bucket D_i is identified by $\mathcal{M}(K) \rightarrow D_i$. 2) The index of D_i is used for determining the correct chunk C_k of the first requested record $R_S = (K_S, V_S)$. 3) Within C_k a binary search is performed for finding R_S. The binary search massively benefits from equally sized records. 4) A sequential read is performed until K_E was found and consequently R_E returned. It might be needed to perform the sequential read over chunk and bucket boundaries.

Single select process A request of a single row (single select) is considered as special case of the range select process where $K_S = K_E$. Therefore, it is covered by step 1 to 3.

Components of DRUMS

BucketContainer and its buckets

The BucketContainer is a buffer that is organized in buckets B (memory buckets). It manages the distribution of incoming records to the buckets in RAM. As in DRUM, the distribution of the incoming records $R = (K, V)$ to the Buckets B is based on a predefined mapping function $\mathcal{M}(K) \rightarrow B_i$.

The BucketContainer is decoupled from any I/O-operation, so that preparing the data for writing can be done in parallel to the I/O-processes. The larger the size of the BucketContainer, the larger are the parts of the data

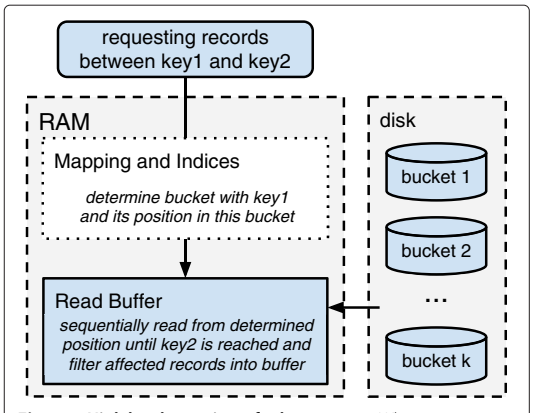


Figure 2 High level overview of select process. When a request is sent to DRUMS, four steps are done to read the requested records. 1) The bucket of interest is determined. 2) The correct chunk of the first requested record is identified, using a sparse index. 3) The position of the requested key-value pair is determined. 4) A sequential read is performed until the requested range is completely processed.

that can be processed sequentially. This increases the performance significantly as sequential I/O-operations are the most efficient on HDDs and SSDs.

SyncManager, SyncProcess, and Synchronizer

The SyncManager manages all SyncProcesses. It observes the BucketContainer and verifies the preconditions for the synchronisation of buckets B with their counterparts on disk D. If these preconditions are fulfilled, the SyncManager instantiates new SyncProcesses. Several SyncProcesses can be run in parallel. In our implementation, a bucket in memory must reach a predefined fill level or age to be synchronized.

A new SyncProcess is always instantiated with the largest bucket in the BucketContainer fulfilling the above mentioned condition. When a new SyncProcess is started, the affected bucket in the BucketContainer is replaced by an empty one. In this way the synchronization process is not blocking further insert operations for this bucket.

The SyncProcess instantiates new Synchronizers. A Synchronizer is in charge of writing data from the bucket B_i in memory to the bucket D_i on disk. All records are sorted in B_i and in D_i. Hence, the Synchronizer is capable of using mergesort for synchronizing the records in memory with those on disk.

Representation and structure of the data

Each persistent bucket is represented by a file on a hard disk. The file is structured into two parts (see Figure 3): (i) the header with meta information and the index structure referencing chunks of a predefined size and (ii) the rest of the file used for the records to store, which are organized in chunks. A sparse index [15] is applied as it is memory efficient and takes advantage of the order of *psDrds*.

Whenever a bucket D is opened for reading or writing, the header and the index are read into memory. In this way, a rapid access to the required chunks is possible.

The internal representation of a record in a chunk is a sequence of bytes. This sequence is composed of a key-part and a value-part. Each part may consist of several subparts, each of its own data-type (e.g., integer, long, char or even high level data structures like objects). Because of the fact that each record is of equal size, data structures and memory can be easily reused by application of the adaptor and the prototype pattern [16].

Implementation of DRUMS

DRUMS is build upon Oracle Java 1.6. Therefore, it is platform independent. We developed DRUMS in an atomic thread-based way. All components work asynchronously and are exchangeable. This allows fast adaptations on single subprocesses or exchanging whole components like the Synchronizer.

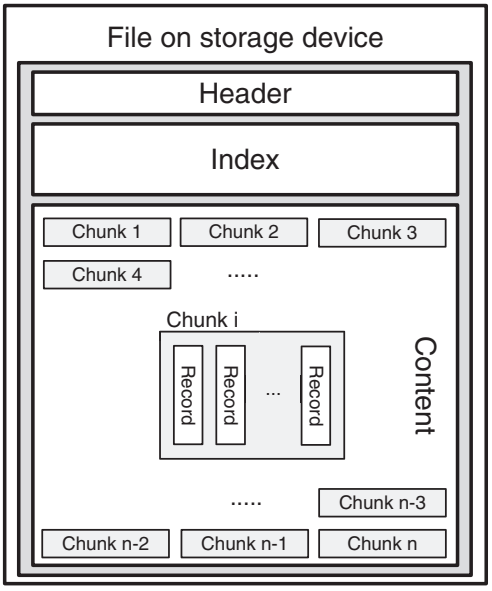


Figure 3 Structure of a file on storage device. The file is structured into (i.) a header, (ii.) an index structure and (iii.) the content, containing the records.

Results and discussion

In this section we first give a short introduction into two different *psDrd* sets used for evaluation. Second, we present the results and the evaulation approach considering (i) inserts, (ii) random lookups, and (iii) random range selects.

To prove the superiority of DRUMS in comparison with standard solutions within a desktop environment, we compare it to MySQL which is used widely in the bioinformatics community.

Two different *psDrd* sets are evaluated. The data sets are described below. DRUMS as well as MySQL were tested comparatively using the three measures: (i) - (iii). For all tests a standard desktop computer was used. MySQL as well as DRUMS are limited to use only 2 GB of the available memory. Details can be obtained from Table 1.

Data sets

SNP-Data from the 1001 genomes project

The 1001 Genomes Project [17,18] has the goal to understand the resulting of small mutations in different accessions of the reference plant Arabidopsis thaliana.

Nettling *et al. BMC Bioinformatics* 2014, **15**:38
http://www.biomedcentral.com/1471-2105/15/38

Table 1 Test system

Processor	Intel Xeon E31225 (4 native cores, no hyperthreading)
Memory	8 GB
Operation system	Debian 6.0 (Squeeze)
Hard drive	Western digital WD10EALX-759, 32 MB cache

The desktop system which was used for the tests. MySQL as well as DRUMS are limited to use only 2 GB of the available memory.

Each accession mainly consists of five attributes: accession identifier, sequence identifier, position on sequence, source base, and target base. We downloaded filtered quality data of the strains sequenced by the Gregor Mendel Institute and the Salk institute on 2012-01-15, containing 251 data sets, with 137, 369, 902 SNPs. From all files, we extracted the data of the following five columns: accession name, chromosome, position on chromosome reference nucleotide, and mutated nucleotide. For the definitions of the used data types and their configuration (e.g., index properties) used in MySQL and DRUMS see Table 2.

All data are public available at http://1001genomes.org/datacenter/.

HERV data

Human endogenous retroviruses (HERVs) have integrated themselves in the human genome millions of years ago. Because of the high number of existing HERV fragments, they are thought to have a regulatory role. To investigate a possible influence of HERVs, it is needed to locate HERV fragments. Therefore, over 7000 known HERV fragments were blasted against the human genome to find new putative HERV-like regions. In the work of Konstantin Kruse [19] all regions with an E-value less than $1e-20$ were accepted as putative HERV-like region. This lead to 802, 710, 938 single records, stored in 20 files with tab-separated data field, with a total size of 50 GB. From these files we used the following seven columns: query id, subject id, query start, query end, subject start, subject end, and E-value. For the definitions of the used data types and their configuration (e.g., index properties) used in MySQL and DRUMS see Table 3.

Table 2 Data types used for SNP data

Column	MySQL properties	DRUMS properties
Accession name	TINY INT, primary key	1 byte, key part 1
Chromosome	SMALL INT, primary key	2 byte, key part 2
Position on chromosome	INT, primary key	4 byte, key part 3
Reference nucleotide	VARCHAR	1 byte, value part 1
Mutated nucleotide	VARCHAR	1 byte, value part 2

Used data types in MySQL and DRUMS for SNP data. All columns being part of the primary key are indexed.

Insert performance

DRUMS must be able to store hundreds of millions of records. Because of this, it is needed to evaluate the insert performance.

To estimate the insert performance, we measure the time for inserting 10^6 records. We obtain 140 time measurements points in case of SNP-Data and 800 for HERV data. Figures 4a and 4b show the insert performance of DRUMS (blue) and MySQL (green). Despite using bulk-requests for inserting the data, it was impossible to insert all 800 million HERV records into the MySQL instance. MySQL inserts about 200 million records in the first week, but Figure 4b shows that the insert performance has dropped to 300 records per second after one week. The insert performance of DRUMS also decreases, but it was able to insert the whole data set within 4.53 hours. At the end of the test, DRUMS was still able to perform more than 20000 inserts per second.

Figure 4a and 4b show that DRUMS has a better insert performance than MySQL on both test datasets. The insert performance of MySQL and of DRUMS decreases with the number of records already inserted. Regarding MySQL one possible explanation is the continuous reorganistation and rewriting of the index.

The insert performance of DRUMS decreases slowly in comparison to MySQL. The reason for this is the decreasing ratio of read- to write-accesses with each round of synchronisation. With other words, DRUMS must read more and more records per new record to write with the growing amount of data already stored on disk. However, DRUMS still inserts more than 20000 records per second at the end of the insert test for HERV data, corresponding to approximately 400 kB per second.

Performance on random lookups

From the view of bioinformatics, single lookups make no sense in both experiments. However, the performance of single-lookups is a significant indicator for the overall performance and the suitability of the implementation of a tool for handling data sets. Moreover, the test may show how close the measured performance to the theoretical hardware limits of the used standard desktop hardware is. Considering the test environment, it is assumed that a random access would take approximately 20 ms. Hence, if no other disk accesses are done, it would be theoretically possible to read 50 records per second.

Figures 5a and 5b show the performance of MySQL and DRUMS, when performing random lookups. Again, DRUMS performs better than MySQL in case of handling our two data sets. Figure 5a implies that DRUMS is able to do 160 times more random lookups than theoretically possible, when accessing SNP data. In comparison,

Nettling *et al. BMC Bioinformatics* 2014, **15**:38
http://www.biomedcentral.com/1471-2105/15/38

Table 3 Data types used for HERV data

Column	MySQL properties	DRUMS properties
Chromosome	TINY INT, primary key	1 byte, key part 1
Start-position on chromosome	INT, primary key	4 byte, key part 2
End-position on chromosome	INT, primary key	4 byte, key part 3
Start-position on HERV	SMALL INT, primary key	2 byte, key part 4
End-position on HERV	SMALL INT, primary key	2 byte, key part 5
Id of referenced HERV	SMALL INT, primary key	2 byte, key part 6
Strand on chromosome	TINY INT, primary key	1 byte, key part 7
E-value	DOUBLE	4 byte, value part 1

Used data types in MySQL and DRUMS for HERV data. All columns being part of the primary key are indexed.

only 20 random lookups per second are performed when accessing HERV data. The reason for this difference are cache structures provided by the operating system and the underlying hardware.

In case of accessing SNP data, the complete data set might be cached by the operating system after approximately 650,000 lookups. Hence, organizing the SNP data as DRUMS structure results in a file size small enough that it could be loaded into memory. Therefore, nearly each request could be answered from the operating systems cache after a warm up. In contrast, the HERV data set is too large to fit into memory, so only a few random lookups could be answered from cache. The increasing performance of MySQL and DRUMS in Figure 5b is also an indication for the use of caches. Figure 5b shows that DRUMS can perform 20 random lookups of theoretically possible 50.

While considering the experimental results of MySQL, the impression is conveyed that the defined index was not used correctly. However, a closer look validates the results as the explicit MySQL index for the SNP table has the size of 2380 MB, which will not fit into the allowed 2 GB of main memory. Hence, even index-based searches in MySQL need several accesses to the hard disk resulting in worse performance. In contrast, the sparse index of each bucket of DRUMS requires just 0.5 MB, which sums

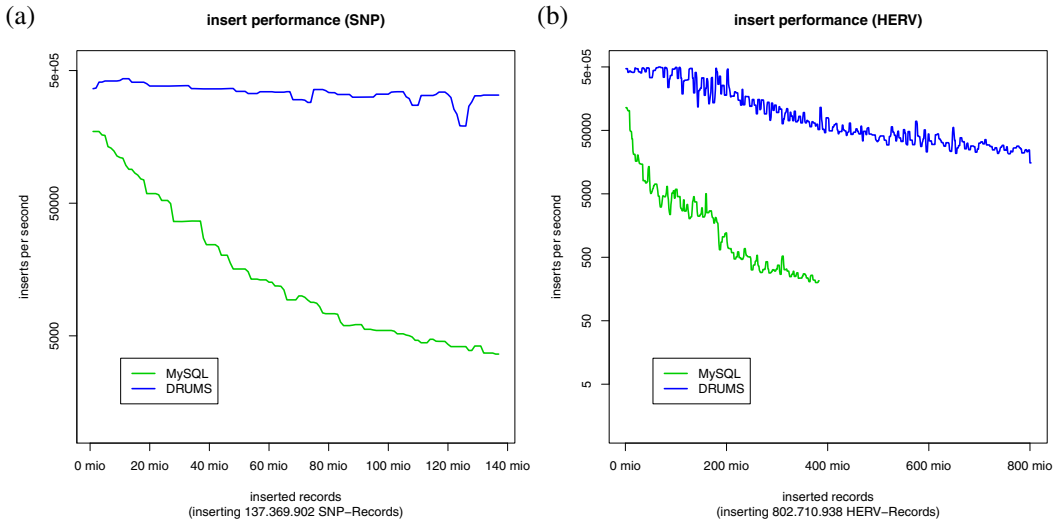


Figure 4 Insert performance. The blue line represents DRUMS, the green line represents MySQL. **(a)** Insert performance on SNP-data **(b)** Insert performance on HERV data. Concerning MySQL, it was impossible to insert all 800 million HERV records. DRUMS inserted the complete data set within 4.53 hours.

Nettling *et al. BMC Bioinformatics* 2014, **15**:38
http://www.biomedcentral.com/1471-2105/15/38

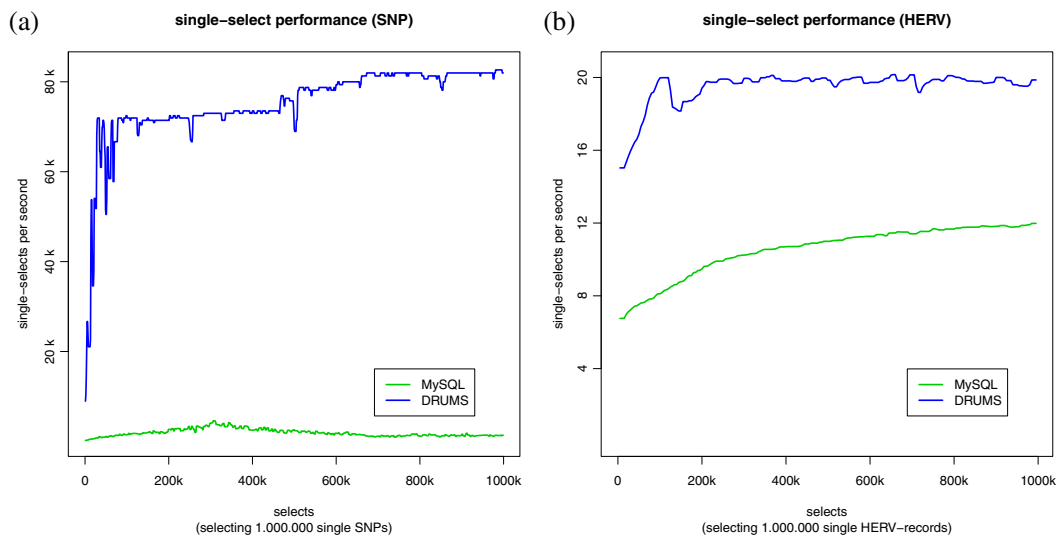


Figure 5 Random lookups performance. The blue line represents DRUMS, the green line represents MySQL. **(a)** Random lookup performance on SNP-data. **(b)** Random lookup performance on HERV data.

up to only 123 MB for all buckets. To find a single record in a chunk, DRUMS performs a binary search. The binary search can be done very efficiently for the reason that all records are of equal size. Because of the reduced demands on the hardware, DRUMS provides a good performance even on very large data sets like HERV.

Performance on random range selects

As described in the section Background, *psDrd*-records are mostly requested by range. Therefore, the need to benchmark the performance of range requests is obvious.

The request for the SNP-data is as follows: Select all SNPs on chromosome c between position x and y for

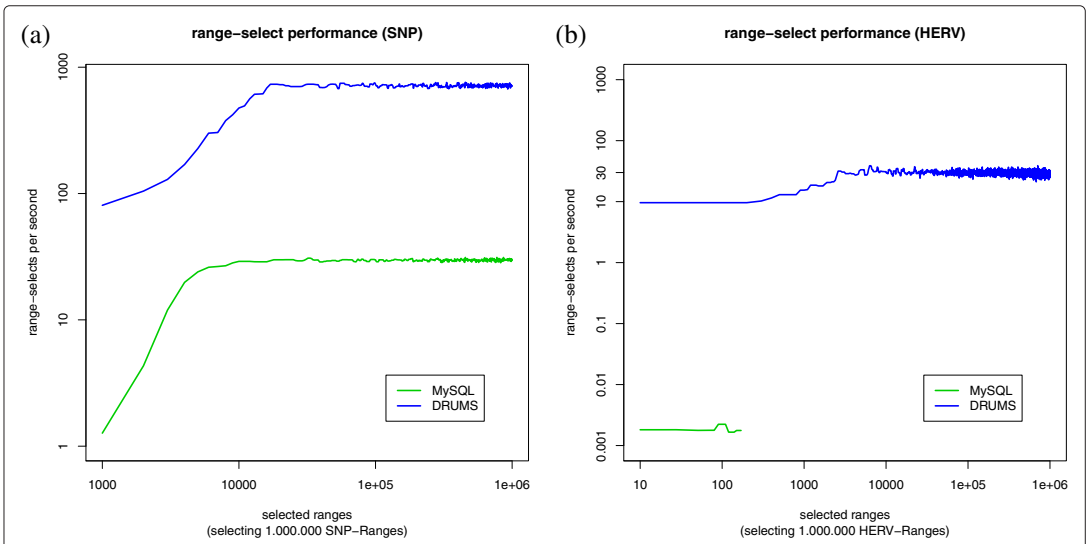


Figure 6 Range select performance. The blue line represents DRUMS, the green line represents MySQL. **(a)** Range select performance on SNP-data. **(b)** Range select performance on HERV data. Concerning MySQL, we stopped the test after 26.35 hours. DRUMS read 64 billion records in 9.61 hours.

all ecotypes in the database. To perform the read test for SNP-data, we first randomly generated 10^6 ranges of length 10^3 to 10^4. Second, we request records within those ranges randomly distributed over the whole genome of Arabidopsis thaliana.

Analogously, we generate 10^6 test requests for the HERV data set with lengths from 10^5 to 10^6. Again, we randomly distributed range-requests over the whole human genome. It might be a common task to filter the requested data by value. MySQL provides this functionality by defining the filter condition in the WHERE-clause. To accomplish this in DRUMS, the returned records must be checked iteratively. In this test, we filter the requested HERV records by an E-value less than 10^{-20}, 10^{-25}, 10^{-30}, 10^{-35}, 10^{-40}, 10^{-45} or 10^{-50}, randomly chosen.

Figures 6a and 6b show the results of the range select test. Once more, both databases perform much better on the smaller SNP-data set. Besides caching, this time another explanation for this observation is that a range request on the SNP-data contains in average 3 times fewer records than a range request on the HERV data. The performance increases with the number of read records. The performance of DRUMS increases by a factor of 10 and of MySQL by a factor of 26. However, DRUMS performs in average on the SNP-data 24 times faster than MySQL.

Regarding the larger HERV data set, DRUMS is able to perform 30 range-selects per second in average. This is over 15000 times faster than MySQL.

Within the whole test, 64 billion records were read in 9.61 hours. That corresponds to an overall read performance of 35.7 MB per second, filtering included. In contrast, MySQL read 6.6 million records in 26.35 hours, which corresponds to only 1.3 kB per second.

Conclusions

We defined *psDrd (position-specific DNA related data)* and showed three important properties of this kind of data. The flaws of DRUM were shown, which is already suitable for storing *psDrd*, but not for requesting it efficiently. The article introduces DRUMS, a data management concept optimized to tackle the challenges of dealing with mid-size data sets in form of *psDrd* using standard desktop technology instead of expensive cluster hardware.

An implementation of the DRUMS concept was compared to the widely spread standard database management solution MySQL considering two data sets of the bioinformatics context. On the larger HERV data set, the evaluated DRUMS implementation was 23 times faster inserting all records, two times faster performing random lookups, and 15456 faster performing range requests. Hence, the experiments show that dealing with *psDrd* benefits significantly from the characteristics of the DRUMS concept. Therefore, our main contribution is suggesting this data management concept for increasing the performance during data intensive processes while keeping the hardware investments low.

Availability and requirements

Project name: DRUMS
Project home page: http://mgledi.github.io/DRUMS
Project home page of examples: http://github.com/mgledi/BioDRUMS
Operating system: Platform independent
Programming language: Java
Other requirements: none
License: GNU GPL v2
Any restrictions to use by non-academics: No specific restrictions.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MN and NT developed and tested the Java code. All of the authors contributed to the design of the software architecture. All of the authors read and approved the final version of the manuscript.

Acknowledgements
We are grateful to Dr. Christiane Lemke and Anika Gross for revising the manuscript. We thank Michael Roeder for testing the installation and usage instructions. Furthermore, we thank *Unister GmbH* for the opportunity to develop and publish the software as open source project.

Author details
[1]Institute of Computer Science, Martin Luther University, Halle (Saale), Germany. [2]R&D, Unister GmbH, Leipzig, Germany. [3]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany.

Received: 31 July 2013 Accepted: 17 January 2014
Published: 4 February 2014

References

1. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11**(8):R86+.
2. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J: **Galaxy: a web-based genome analysis tool for experimentalists.** Current protocols in molecular biology/edited by Frederick M. Ausubel ... [et al.] 2010:Chapter 19.
3. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A: **Galaxy: a platform for interactive large-scale genome analysis.** *Genome Res* 2005, **15**(10):1451–1455.
4. **Single nucleotide polymorphism.** 2012. [http://en.wikipedia.org/wiki/Single_Nucleotide_Polymorphism]
5. Bulyk M: **Computational prediction of transcription-factor binding site locations.** *Genome Biol* 2003, **5:**201+.
6. Nguyen T, Androulakis I: **Recent advances in the computational discovery of transcription factor binding sites.** *Algorithms* 2009, **2:**582–605.
7. Malone J, Oliver B: **Microarrays, deep sequencing and the true measure of the transcriptome.** *BMC Biol* 2011, **9:**34+.
8. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10:**57–63.
9. de Leeuw N, Hehir-Kwa JY, Simons A, Geurts van Kessel A, Smeets DF, Faas BH, Pfundt R: **SNP array analysis in constitutional and cancer genome diagnostics–copy number variants, genotyping and quality control.** *Cytogenet Genome Res* 2011, **135:**212–221.

10. Kihara D, Yang YDD, Hawkins T: **Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools.** *Cancer Inform* 2006, **2:**25–35.
11. Roukos DH: *Next-Generation Sequencing & Molecular Diagnostics*. London: Future Medicine Ltd; 2013.
12. **MySQL classic edition.** 2012. [http://www.mysql.com/products/classic/]
13. **Common wrong data types.** 2012. [http://code.openark.org/blog/mysql/common-data-types-errors-compilation]
14. Lee HT, Leonard D, Wang X, Loguinov D: **IRLbot: scaling to 6 billion pages and beyond.** In *Proceedings of the 17th international conference on World Wide Web, WWW '08*. New York, NY, USA: ACM; 2008:427–436.
15. **Database index - sparse index.** 2012. [http://en.wikipedia.org/wiki/Database_index#Sparse_index]
16. Gamma E, Helm R, Johnson R, Vlissides J: *Design patterns: elements of reusable object-oriented software*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.; 1995.
17. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, Smith LM, Cao J, Fitz J, Warthmann N, Henz SR, Huson DH, Weigel D: **Reference-guided assembly of four diverse Arabidopsis thaliana genomes.** *Proc Nat Acad Sci USA* 2011, **108**(25):10249–10254.
18. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Müller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D: **Whole-genome sequencing of multiple Arabidopsis thaliana populations.** *Nat Genet* 2011, **43**(10):956–963.
19. Kruse K: **Analysis of gene expression in correlation to endogenous retroviruses.** Martin Luther University, Halle (Saale) Germany 2011. [Bachelor Thesis]

doi:10.1186/1471-2105-15-38
Cite this article as: Nettling *et al.*: DRUMS: Disk Repository with Update Management and Select option for high throughput sequencing data. *BMC Bioinformatics* 2014 **15**:38.

5 Predicting transcription factor binding sites using Phylogenetic Footprinting

5.1 Detecting and correcting the binding-affinity bias in ChIP-Seq data using inter-species information

M Nettling, H Treutler, J Cerquides, I Grosse. 2016. Detecting and correcting the binding-affinity bias in ChIP-Seq data using inter-species information. *BMC Genomics* 17:1. *doi:10.1186/s12864-016-2682-6*

Nettling *et al. BMC Genomics* (2016) 17:347
DOI 10.1186/s12864-016-2682-6

BMC Genomics

METHODOLOGY ARTICLE Open Access

Detecting and correcting the binding-affinity bias in ChIP-seq data using inter-species information

Martin Nettling[1*], Hendrik Treutler[2], Jesus Cerquides[3] and Ivo Grosse[1,4]

Abstract

Background: Transcriptional gene regulation is a fundamental process in nature, and the experimental and computational investigation of DNA binding motifs and their binding sites is a prerequisite for elucidating this process. ChIP-seq has become the major technology to uncover genomic regions containing those binding sites, but motifs predicted by traditional computational approaches using these data are distorted by a ubiquitous binding-affinity bias. Here, we present an approach for detecting and correcting this bias using inter-species information.

Results: We find that the binding-affinity bias caused by the ChIP-seq experiment in the reference species is stronger than the indirect binding-affinity bias in orthologous regions from phylogenetically related species. We use this difference to develop a phylogenetic footprinting model that is capable of detecting and correcting the binding-affinity bias. We find that this model improves motif prediction and that the corrected motifs are typically softer than those predicted by traditional approaches.

Conclusions: These findings indicate that motifs published in databases and in the literature are artificially sharpened compared to the native motifs. These findings also indicate that our current understanding of transcriptional gene regulation might be blurred, but that it is possible to advance this understanding by taking into account inter-species information available today and even more in the future.

Keywords: Binding-affinity bias, ChIP-seq, Phylogenetic footprinting, Evolution, Transcription factor binding sites, Gene regulation

Background

Predicting transcription factor binding sites and their motifs is essential for understanding transcriptional gene regulation and thus of importance in almost all areas of modern biology, medicine, and biodiversity research [1, 2]. Countless approaches exist for predicting motifs from these genomic regions [3–6], but predicting motifs from ChIP-seq data and similar experimental data is hampered by the contamination with false positive genomic regions as well as the enrichment of high-affinity binding sites [7–9].

The contamination with false positive genomic regions is caused by at least three reasons. First, the transcription factor or other DNA binding protein pulled down by immunoprecipitation may not bind directly to the binding site [10]. Second, ChIP-seq target regions may not contain a binding site due to experimental settings such as sequencing depth or DNA fragment length [11, 12]. Third, false positive regions may be predicted in the subsequent ChIP-seq data analysis due to never perfect analysis pipelines and too low signal cutoff thresholds [8]. These three effects may lead to the selection of false positive ChIP-seq regions that do not contain at least one binding site.

The enrichment of high-affinity binding sites is caused by at least two reasons. First, most antibodies have a preference of binding high-affinity binding sites with a higher probability than low-affinity binding sites, causing the set

*Correspondence: martin.nettling@informatik.uni-halle.de
[1]Institute of Computer Science, Martin Luther University, Halle (Saale), Germany
Full list of author information is available at the end of the article

of binding sites bound in the ChIP-seq experiment to be partially different from the set of binding sites bound in vivo [13, 14]. Second, true positive regions with low-affinity binding sites are rejected due to too high signal cutoff thresholds [5, 8]. These two effects may lead to an under-representation of low-affinity binding sites and an over-representation of high-affinity binding sites in ChIP-seq regions.

Taken together, the contamination with false positive genomic regions leads to the *contamination bias* [15] and thus to the prediction of artificially softened motifs, whereas the enrichment of sequences with high-affinity binding sites leads to the *binding-affinity bias* [16] and thus to the prediction of artificially sharpened motifs. Neglecting these effects leads to distorted motifs and could potentially affect all downstream analyses [17–20]. Existing approaches for predicting motifs are capable of detecting and correcting the contamination bias, which has been found to increase the quality of motif prediction considerably [8, 21, 22], and here we investigate the possibility of detecting and correcting the binding-affinity bias.

Detecting the binding-affinity bias seems impossible based on sequence data from one species alone, but it seems possible based on inter-species information. This is possible due to the fact that the binding-affinity bias is stronger in the target regions of the ChIP-seq experiment in the reference species than in orthologous regions of phylogenetically related species. This stronger binding-affinity bias yields more biased motifs in the reference species than in phylogenetically related species, and this difference may be used for detecting and potentially correcting the binding-affinity bias.

Phylogenetic footprinting models typically (i) take into account ChIP-seq data of only one species and (ii) do not take into account heterogeneous substitution rates among different DNA regions, heterotachious evolution of DNA regions, and loss-of-function mutations in binding sites. The consideration of (i) ChIP-seq data of more than one species and (ii) heterogeneity, heterotachy, and loss-of-function mutations are likely to improve both phylogenetic footprinting as well as the detection and correction of the binding-affinity bias, but in this work we investigate if the detection and correction of this bias is possible based on (i) ChIP-seq data of only one species and (ii) a simple phylogenetic footprinting model that neglects heterogeneity, heterotachy, and loss-of-function mutations.

We first investigate if the effect of observing more biased motifs in the reference species than in phylogenetically related species is measurable beyond statistical noise in target regions of five ChIP-seq data sets of human and in orthologous regions of monkey, dog, cow, and horse. We then develop a phylogenetic footprinting model that incorporates the binding-affinity bias, investigate if this model improves or deteriorates motif prediction compared to traditional models that do not incorporate it, and compare the motifs predicted with and without the correction of the binding-affinity bias.

Results and discussion

In subsection "Using sequence-information of phylogenetically related species to detect the binding-affinity bias", we describe the basic idea of how the binding-affinity bias could be detected based on inter-species information using a toy example. In the remaining subsections we perform three studies based on ChIP-seq data sets of five transcription factors and on multiple alignments of the human ChIP-seq target regions with orthologous regions from monkey, dog, cow, and horse. In subsection "Decrease of information contents in motifs from phylogenetically related species" we investigate if the effect of observing more biased motifs in the reference species than in phylogenetically related species is measurable in these five data sets. In subsection "Modeling the binding-affinity bias increases classification performance", we investigate if a correction of the binding-affinity bias leads to an improvement or a deterioration of the classification performance. In subsection "Modeling the binding-affinity bias leads to softened motifs", we compare the sequence motifs predicted with and without the correction of the binding-affinity bias.

Using sequence-information of phylogenetically related species to detect the binding-affinity bias

Detecting and correcting the binding-affinity bias might be possible because the binding-affinity bias inherent to the ChIP-seq experiment in the reference species (Fig. 1a) is stronger than the indirect binding-affinity bias in orthologous regions from phylogenetically related species. Under this assumption, the information content of the predicted motifs [23] should decrease with the phylogenetic distance from the reference species due to the increasing number of mutations.

To illustrate this idea, we present a toy example consisting of six binding sites from four phylogenetically related species in Fig. 1b and Table 1. In this toy example, we assume an exaggerated binding-affinity bias of three high-affinity binding sites captured by the ChIP-seq experiment and three low-affinity binding sites not captured by the ChIP-seq experiment. In real world applications the native motif is unknown and the motif predicted on the available data is biased to an unknown degree. In the presented toy example, however, the native motif is considered to be known so that the effect of the binding-affinity bias on the motifs of the reference species (species 1) and the phylogenetically related species (species 2, 3, and 4) can be illustrated.

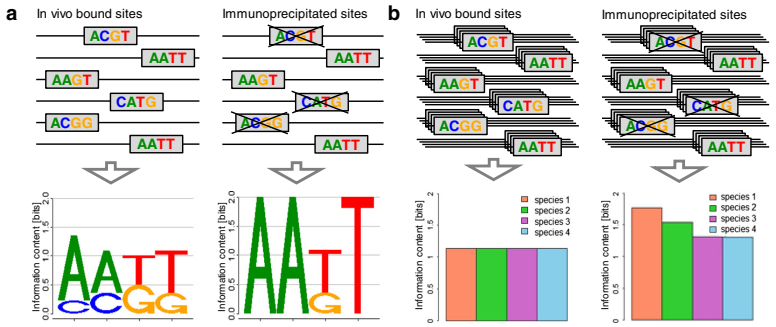


Fig. 1 Influence of binding-affinity bias on information content. **a** Binding-affinity bias in the reference species. The left column shows binding sites bound in vivo as well as the sequence logo. In the right column, enrichment of high-affinity binding sites by chromatin immunoprecipitation leads to a different motif with higher information content. **b** Binding-affinity bias in the reference species and three phylogenetically related species. The left column shows binding sites bound in vivo and the information content of the species–specific motifs. In the right column, the enrichment of high-affinity binding sites in the reference species and the other three species leads to different motifs with different information content in each species. The effect of this enrichment decreases with the phylogenetic distance from the reference species as reflected by decreasing information contents. Please find the sequences of all species in Table 1

Table 1 Influence of binding-affinity bias on information content. We illustrate the effect of binding-affinity bias with the given toy example of a ChIP-seq experiment for six binding sites in four species. Due to low binding-affinity, red binding sites are insufficiently bound. This results in the absence of red binding sites in the measured data which we denote binding-affinity bias. Binding sites with low binding-affinity typically comprise dissimilar bases in contrast to black binding sites with high affinity and common bases. The absence of red binding sites leads to a sharpening of the resulting motif, which we indicate using the information content. The information content without binding-affinity bias is equal in all species, whereas the information content with binding-affinity bias increases in all species. The vital point is that the effect of binding-affinity bias decreases with phylogenetic distance, which involves an increasing number of mutations. Please find a visualization of this toy example in Fig. 1b

	Species 1	Species 2	Species 3	Species 4
Binding site 1	A C G T	A C G T	A C T T	A A T T
Binding site 2	A A T T	A A T T	C A G T	A C G T
Binding site 3	A A G T	C A T G	A A G T	A A T G
Binding site 4	C A T G	A A G T	A C T G	A A G T
Binding site 5	A C G G	A C G G	A A G T	C A G T
Binding site 6	A A T T	A A T T	A A T G	A C T G
Number of mutations in all binding sites	0	6	9	14
Information content without binding-affinity bias	1.13	1.13	1.13	1.13
Information content with binding-affinity bias	1.77	1.54	1.31	1.31

The motif predicted from the three target regions containing high-affinity binding sites is strongly biased in reference species 1, and it is impossible to predict the native motif from only those three target regions. However, a shadow of this strong binding-affinity bias also exists in orthologous regions of species 2, 3, and 4, so the motifs predicted from these orthologous regions in species 2, 3, and 4 are biased, too. This bias in species 2, 3, and 4, however, is weaker than the bias in reference species 1, and this difference can be exploited for detecting and correcting the binding-affinity bias and for predicting the native motif from the three target regions of high-affinity binding sites in reference species 1 and their orthologous regions in species 2, 3, and 4.

Specifically, the binding-affinity bias introduced by the ChIP-seq experiment in reference species 1 causes a strong increase of the information content of the predicted motif (1.77 bit) compared to the native motif (1.13 bit). The shadow of the binding-affinity bias in species 2, 3, and 4 also causes an increase of the information contents of the motifs predicted in species 2 (1.54 bit), species 3 (1.31 bit), and species 4 (1.31 bit), but this increase in species 2, 3, and 4 is smaller than in reference species 1 (Table 1 and Fig. 1b). The increase of information content decreases with the number of observed mutations and thus the phylogenetic distance of species 2, 3, and 4 to reference species 1 in which the ChIP-seq experiment has been performed. Hence, the observation of a decreased information content of motifs predicted in orthologous regions of phylogenetically related species compared to the information content of the motif predicted in the

reference species could indicate the presence of a binding-affinity bias and possibly allow the correction of that bias.

Decrease of information contents in motifs from phylogenetically related species

We investigate this hypothesis on human ChIP-seq data of five transcription factors [10, 24] and multiple alignments of the human ChIP-seq target regions with orthologous regions from monkey, dog, cow, and horse [25] ("Data" Methods). We calculate the information contents of motifs from human (reference species), monkey, dog, cow, and horse for each of the five data sets ("Decrease of information contents in motifs from related species" Methods) and present the results in Fig. 2. We find for each of the five data sets that the information content of the motif from the reference species is significantly higher ($p < 1.83 \times 10^{-14}$, Wilcoxon Signed-Rank Test, Additional file 1: Table S1) compared to the information contents of the motifs from monkey, dog, cow, and horse.

Modeling the binding-affinity bias increases classification performance

Motivated by this observation, we develop a phylogenetic footprinting model capable of taking into account the contamination bias ($\mathcal{M}^{\mathrm{C}}_{-}$), the binding-affinity bias ($\mathcal{M}^{-}_{\mathrm{BA}}$), neither one or the other $\mathcal{M}^{-}_{-}$, or both ($\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$) ("Modeling the binding-affinity bias" Methods and Additional file 1: Section 1). In order to study to which degree these models are capable of modeling multiple alignments originating from ChIP-seq data, we consider the principle of parsimony [26], which states that the simplest of competing explanations is the most likely to be correct. As the new model $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ is more complex than the traditional model $\mathcal{M}^{\mathrm{C}}_{-}$, we should accept it only if it provides a more accurate representation of the data. A standard approach for measuring how accurately a model represents a data set is to measure its performance of classifying, in this case, motif-bearing and non-motif-bearing alignments, and a standard approach for measuring classification performance is stratified repeated random sub-sampling validation ("Measuring classification performance" Methods, Fig. 5).

Using this approach we measure the performance of the four models $\mathcal{M}^{-}_{-}$, $\mathcal{M}^{-}_{\mathrm{BA}}$, $\mathcal{M}^{\mathrm{C}}_{-}$, and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ to classify each of the five data sets against the other four. Fig. 3a shows that $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ yields a higher classification performance than $\mathcal{M}^{\mathrm{C}}_{-}$ in all five data sets ($p < 2.3 \times 10^{-17}$, Wilcoxon Signed-Rank Test, Additional file 1: Table S2), indicating that the new model $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ is more realistic than the traditional model $\mathcal{M}^{\mathrm{C}}_{-}$. We also find that $\mathcal{M}^{-}_{\mathrm{BA}}$ yields a significantly higher classification performance than $\mathcal{M}^{\mathrm{C}}_{-}$ in all five data sets ($p < 1.8 \times 10^{-17}$, Wilcoxon Signed-Rank Test), which indicates that taking into account the binding-affinity bias has a larger impact on the classification performance than taking into account the contamination bias (Additional file 1: Figure S1, Figure S2, Figure S10, Figure S11, Figure S12, Figure S13, Figure S14, Figure S15 and Figure S16).

Modeling the binding-affinity bias leads to softened motifs

Next, we investigate the information contents of the corrected motifs predicted by models $\mathcal{M}^{-}_{\mathrm{BA}}$ and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ that take into account the binding-affinity bias and the traditional motifs predicted by models $\mathcal{M}^{-}_{-}$ and $\mathcal{M}^{\mathrm{C}}_{-}$ that neglect this bias. Fig. 3b shows that the information contents of motifs predicted by $\mathcal{M}^{\mathrm{C}}_{-}$ are significantly higher than the information contents of motifs predicted by $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ ($p < 4.0 \times 10^{-18}$, Wilcoxon Signed-Rank Test). We also find that the information contents of motifs predicted by $\mathcal{M}^{-}_{-}$ are higher than the information contents of motifs predicted by $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ ($p < 4.0 \times 10^{-18}$, Wilcoxon Signed-Rank Test, Additional file 1: Table S4), stating that

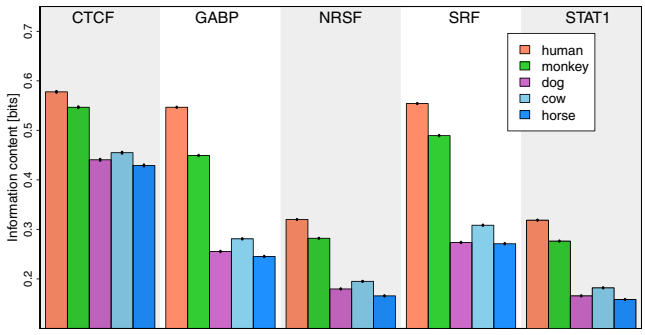


Fig. 2 Mean information content and standard error for motifs of five transcription factors in five species. The information content of motifs in the reference species (human) is significantly higher compared to the four phylogenetically related species ($p < 1.8 \times 10^{-14}$). The information content typically decreases with the phylogenetic distance from the reference species

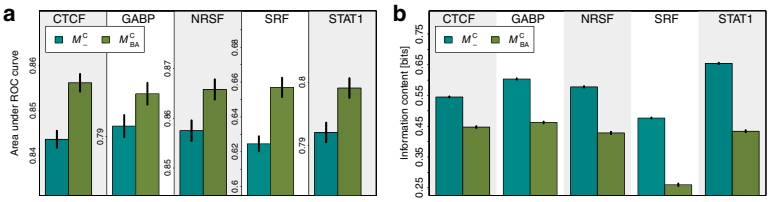


Fig. 3 Comparison of models $\mathcal{M}^{\mathrm{C}}_{-}$ and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$. **a** Mean classification performance and standard error of the two models $\mathcal{M}^{\mathrm{C}}_{-}$ and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ quantified by the area under the receiver operating characteristic curve. We find for each of the five data sets a significantly increased classification performance for $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ compared to $\mathcal{M}^{\mathrm{C}}_{-}$. Examples for ROC curves are shown in Additional file 1: Figure S10, Figure S11, Figure S12, Figure S13, Figure S14 and Figure S15. **b** Mean information content and standard error of the motifs predicted by the two models $\mathcal{M}^{\mathrm{C}}_{-}$ and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$. We find for each of the five data sets a significantly decreased information content in motifs predicted by $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ compared to $\mathcal{M}^{\mathrm{C}}_{-}$ ($p < 4.0 \times 10^{-18}$)

the binding-affinity bias is stronger than the contamination bias. Equivalently, this states that the joint effect of both biases leads to an artificial sharpening of the motifs and an artificial overestimation of the binding affinities (Additional file 1: Figure S3, Figure S4, Figure S17, Figure S18).

Finally, we inspect the differences of the corrected motifs predicted by $\mathcal{M}^{-}_{\mathrm{BA}}$ and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ and the traditional motifs predicted by $\mathcal{M}^{-}$ and $\mathcal{M}^{\mathrm{C}}_{-}$. Fig. 4 shows the differences between the base distributions of pairs of motifs for $\mathcal{M}^{\mathrm{C}}_{-}$ and $\mathcal{M}^{\mathrm{C}}_{\mathrm{BA}}$ by difference logos ("Visualizing motif differences with DiffLogo" Methods). We find for each of the five data sets that the corrected motifs are softer than the traditional motifs distorted by the binding-affinity bias. Specifically, we find that the amount of decrease of the most abundant bases in the corrected motifs compared to the traditional motifs is roughly proportional to the base abundance, whereas the increase of the remaining bases is not proportional to the base abundance. Hence, the corrected motifs are not simply a uniformly softened version of the traditional motifs, but motifs with different degrees of dissimilarity at different positions (Additional file 1: Figure S5, Figure S6,Figure S7, Figure S8 and Figure S9).

Conclusions

We studied the possibility of detecting and correcting the binding-affinity bias in ChIP-seq data using interspecies information. We found that the fact that this bias is stronger in target regions of the reference species than its shadow in orthologous regions of phylogenetically related species enables the detection and correction of this bias. We proposed a phylogenetic footprinting model capable of taking into account the binding-affinity bias in addition to the contamination bias, and we applied this model and its three special cases that neglect one of the two biases or both to five ChIP-seq data sets. We found by stratified repeated random sub-sampling validation that taking into account the binding-affinity bias always improves motif prediction, that the motif binding-affinity bias leads to a distortion of motifs that is even stronger than the distortion caused by the contamination bias, and that the corrected motifs are typically softer than those predicted by traditional approaches. The comparison of corrected and traditional motifs showed small but noteworthy differences, suggesting that the refinement of traditional motifs from databases and from the literature might lead to the prediction of novel binding sites, *cis*-regulatory modules, or gene-regulatory networks and might thus advance our attempt of understanding transcriptional gene regulation as a whole.

Methods

In this section we describe "Decrease of information contents in motifs from related species" (i) the determination of the information contents of motifs in the reference species and phylogenetically related species, "Modeling the binding-affinity bias" (ii) the phylogenetic footprinting model that can take into account the binding-affinity bias, the contamination bias, neither one or the other, or both, "Measuring classification performance" (iii) the measurement of the classification performance of these four phylogenetic footprinting models using stratified repeated random sub-sampling validation, and "Visualizing motif differences with DiffLogo" (iv) the visualisation of differences between the corrected and the traditional motifs.

Decrease of information contents in motifs from related species

We determine the information content $I(P)$ of a motif P as described in [23]:

$$H_\ell(P) = \log_2(|\mathcal{A}|) - \sum_{a\in\mathcal{A}} p_{\ell,a} \cdot \log_2(p_{\ell,a})$$
$$I(P) = \sum_{\ell=1}^{W} H_\ell(P), \qquad (1)$$

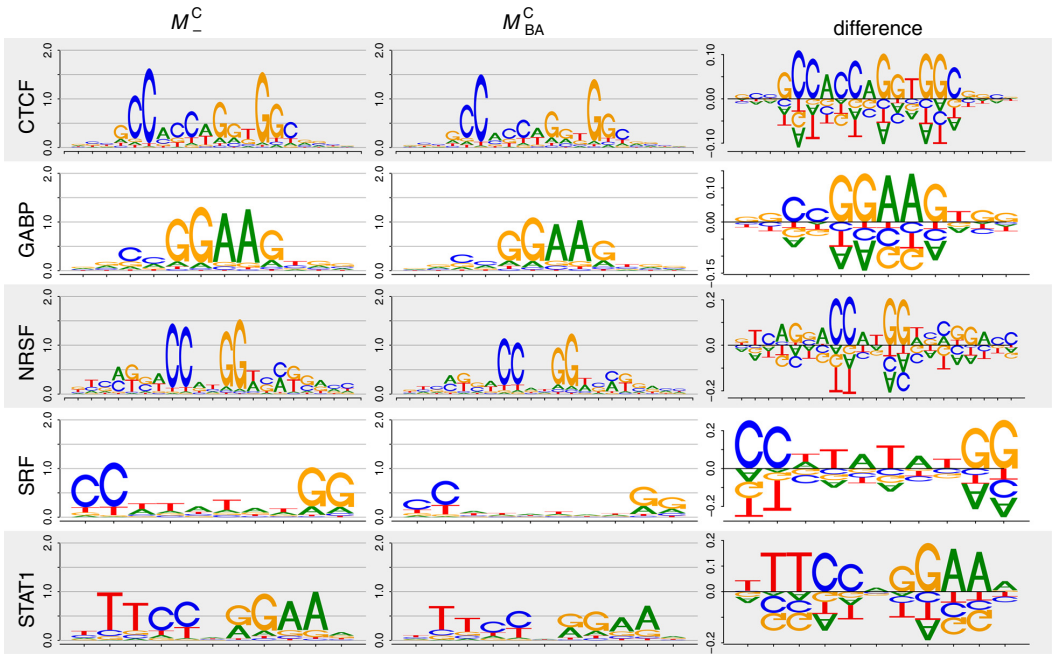


Fig. 4 Differences of motifs predicted by $\mathcal{M}^{C}_{-}$ and $\mathcal{M}^{C}_{BA}$. The height of the base stacks is quantified by the Jensen-Shannon divergence, where high base stacks represent high motif differences. We find significant motif differences exceeding 0.1 bit for all five data sets (Additional file 1: Figure S5, Figure S6, Figure S7, Figure S8 and Figure S9)

where $\mathcal{A} = A, C, G, T$ is the alphabet, $p_{\ell,a}$ is the probability of base a at position ℓ in motif P, and $H_\ell(P)$ denotes the information content of position ℓ in motif P.

We measure the information contents of motifs in five species using repeated random sub-sampling as follows. Initially, we choose one motif for each of the transcription factors CTCF, GABP, NRSF, SRF, and STAT1 from the JASPAR database, namely MA0139.1 for CTCF, MA0062.2 for GABP, MA0138.2 for NRSF, MA0083.2 for SRF, and MA0137.3 for STAT1 [27]. In the first step, we generate a test set from the set of positive alignments (Table 2) by removing randomly 200 alignments. In the second step, we predict for each transcription factor one binding site per target region in all target regions of the reference species (human) in the corresponding test data set, extract the predicted binding sites from the reference species as well as the binding sites at the same positions in the orthologous regions, and calculate for each species the information content of the resulting motif as specified above. We perform both steps 100 times and report the mean and standard error of the information content for each of the five species.

Modeling the binding-affinity bias

In this section we describe the probabilistic model for modeling the binding-affinity bias as a data generating process. A derivation of the log-likelihood for motif-bearing and non-motif-bearing alignments can be found in Additional file 1: Section 1.

Let O be the number of species. A data set comprises N independent multiple sequence alignments. We use X_n to refer to the n-th sequence alignment. Every alignment is formed by O sequences. The o-th

Table 2 Data set statistics for human ChIP-seq data. For each of the five transcription factors (TFs) CTCF, GABP, NRSF, SRF, and STAT1, we specify the (i) average length of transcription factor binding site (TFBS), the (ii) number of alignments, and the (iii) average length of alignments

TF	TFBS length	Number of alignments	Avg. length
CTCF	20 bp	467	213 bp
GABP	12 bp	451	236 bp
NRSF	21 bp	460	245 bp
SRF	12 bp	394	242 bp
STAT1	11 bp	360	244 bp

sequence is denoted by $X_n^{\cdot,o}$. By convention, the reference species (that in which the selection process has taken place) is species 1. Each sequence of alignment X_n is composed of L_n nucleotides. We denote by $X_n^{u,o}$ the u-th nucleotide of the o-th sequence of the n-th alignment. All nucleotides are presented by the set $\mathcal{A} = \{A, C, G, T\}$.

We assume the existence of a common ancestor of all of O species. The sequence of the common ancestor of the n-th alignment is a hidden variable Y_n, with Y_n^u representing its u-th nucleotide. The substitution probability that nucleotide Y_n^u is substituted by the nucleotide $X_n^{u,o}$ is denoted by the variable γ_o.

An alignment X_n may contain a binding site or not. This is denoted by the variable M_n. The length of the binding site is denoted by the variable W and the position of the binding site in alignment X_n is denoted by the variable ℓ_n.

The n-th alignment X_n is sampled as follows. The first decision to be made is whether or not the alignment contains a binding site. This is denoted by variable M_n which follows a Bernoulli distribution with parameter $1 - \alpha$. Thus, whenever variable M_n is equal to 1 (M_n^1), the alignment contains a binding site and when M_n is equal to 0 (M_n^0), it does not.

Thus, parameter α is the probability that alignment X_n contains no binding site. If α equals 0, the sampled data is uncontaminated, because all alignments contain a copy of the binding site. The larger the value of α, the higher the percentage of non motif-bearing alignments in the sampled data. A value of α equal to 1 models a data set where no binding sites are present.

Next we introduce the data generating process for non-motif-bearing alignments and later we explain that for motif-bearing alignments.

1. Sample the primordial sequence as follows: For each position u of the sequence sample nucleotide Y_n^u from the background equilibrium distribution π_0 independent of the previous nucleotides.
2. For each of the descent species $o \in \{1, \ldots, O\}$, sample its sequence given the primordial sequence as follows: To sample nucleotide u of the descent species o, we apply to nucleotide u of the primordial sequence the F81 [28] mutation model with the background equilibrium distribution π_0 and the substitution probability γ_o.

The generating process for motif-bearing sequences is slightly more complex, since it has to deal both with the generation of the binding site and with the selection process. First, we describe how to sample an alignment without taking into account the selection process. Second, we show how to modify this procedure so that the selection process is considered.

Sample a motif-bearing alignment X_n as follows:

1. Sample the start position of the binding site ℓ_n from the uniform distribution.
2. Sample the primordial sequence. For each position u of the sequence outside the binding site, we sample nucleotide Y_n^u from the background equilibrium distribution π_0. For each position u of the binding site, we sample nucleotide Y_n^u from the equilibrium distribution $\pi_{u-\ell_n+1}$.
3. For each of the descent species $o \in \{1, \ldots, O\}$, sample its sequence $X_n^{\cdot,o}$ as follows: For each position u of the descent species o outside the binding site, apply to nucleotide $X_n^{u,o}$ of the primordial sequence the F81 mutation model taking as equilibrium distribution π_0. For each position u of the descent species o inside the binding site, apply to nucleotide $X_n^{u,o}$ of the primordial sequence the F81 mutation model taking as equilibrium distribution $\pi_{u-\ell_n+1}$.

Finally, to model the selection process, we introduce the variable β. β is used to quantify the degree of the binding-affinity bias in the reference species. We assume that a transcription factor binds binding site B with a probability proportional to $p(B|\pi)^{\beta-1}$. As B occurs in vivo with probability $p(B|\pi)$, it occurs in the set of immunoprecipitated sequences with a probability proportional to $p(B|\pi) \cdot p(B|\pi)^{\beta-1} = p(B|\pi)^{\beta}$.

We can interpret the meaning of β as follows: If β is greater than one, low-affinity binding sites are more frequently rejected with respect to $p(B)$ and high-affinity binding sites are less frequently rejected with respect to $p(B)$. This leads to an under-representation of low-affinity binding sites and an over-representation of high-affinity binding sites in the ChIP-seq data set, thus modeling a data set that is affected by the binding-affinity bias. If β is equal to one, low-affinity binding sites are rejected as frequently as high-affinity binding sites, leading to a representative set of binding sites in the ChIP-seq data set, which is not affected by the binding-affinity bias.

Based on that selection model, sample a motif-bearing alignment that has passed the selection process as follows:

1. Sample a motif-bearing alignment disregarding the selection process following the procedure specified above.
2. Decide whether the alignment is accepted or rejected based on the probability of acceptance of the binding site found at the reference species. If the alignment is rejected, go to step 1.

Thus, we denote (i) the model with $\alpha = 0$ and $\beta = 1$ by $\mathcal{M}_{-}^{-}$, (ii) the model with with $\alpha > 0$ and $\beta = 1$ by

$\mathcal{M}_{-}^{C}$, (iii) the model with $\alpha = 0$ and $\beta > 1$ by $\mathcal{M}_{BA}^{-}$, and (iv) the model with $\alpha > 0$ and $\beta > 1$ $\mathcal{M}_{BA}^{C}$. $\mathcal{M}_{-}^{-}$ can neither handle the contamination bias nor the binding-affinity bias. $\mathcal{M}_{-}^{C}$ can only handle the contamination bias, but not the binding-affinity bias. $\mathcal{M}_{BA}^{-}$ can only handle the binding-affinity bias, but not the contamination bias. And $\mathcal{M}_{BA}^{C}$ can handle both the contamination bias and the binding-affinity bias.

We call $\mathcal{M}_{-}^{-}$, $\mathcal{M}_{-}^{C}$, $\mathcal{M}_{BA}^{-}$, and $\mathcal{M}_{BA}^{C}$ *foreground models*. For modeling the background alignments, we use the model with $\alpha = 1$ and $\beta = 1$, which we call *background model* and which we denote by $\mathcal{B}$.

Measuring classification performance

For measuring the classification performance of the four models $\mathcal{M}_{-}^{-}$, $\mathcal{M}_{BA}^{-}$, $\mathcal{M}_{-}^{C}$, and $\mathcal{M}_{BA}^{C}$ we perform stratified repeated random sub-sampling validation as illustrated in Fig. 5 using data sets of the five human transcription factors CTCF, GABP, NRSF, SRF, and STAT1 that have been used for benchmarking the phylogenetic footprinting program *MotEvo* [25].

In step 1, we generate two training sets and two disjoint test sets for each of the five transcription factors as follows. We randomly select 200 alignments from the set of alignments (Table 2) of a particular transcription factor as positive training set, and we choose the set of the remaining alignments as positive test set. We randomly select 500 alignments from the set of alignments of the four remaining transcription factors as negative training set and another disjoint set of 500 alignments as negative test set.

In step 2, we train a foreground model ($\mathcal{M}_{-}^{-}$, $\mathcal{M}_{BA}^{-}$, $\mathcal{M}_{-}^{C}$, or $\mathcal{M}_{BA}^{C}$) on the positive training set and a background model ($\mathcal{B}$) on the negative training set by expectation maximization [29] using a numerical optimization procedure in the maximization step.

We restart the expectation maximization algorithm, which is deterministic for a given data set and a given initialization, 150 times with different initializations and choose the foreground model and the background model with the maximum likelihood on the positive training data and the negative training data, respectively, for classification. We use a likelihood-ratio classifier of the two chosen foreground and background models, apply this classifier to the disjoint positive and negative test sets, and calculate the receiver operating characteristics curve, the precision recall curve, and the area under both curves as measures of classification performance.

We repeat both steps 100 times and determine (i) the mean area under the receiver operating characteristic curve and its standard error and (ii) the mean area under the precision recall curve and its standard error.

Data

The data used in this work originate from human ChIP-seq data of the five human transcription factors CTCF, GABP, NRSF, SRF, and STAT1, where the ChIP-seq data for GABP and SRF published in [10] are available from the QuEST web page [30], and the ChIP-seq data for CTCF, NRSF, and STAT1 published in [24] are available from the SISSRs web page [31]. All five data sets have been filtered for high-quality reads and mapped to a reference

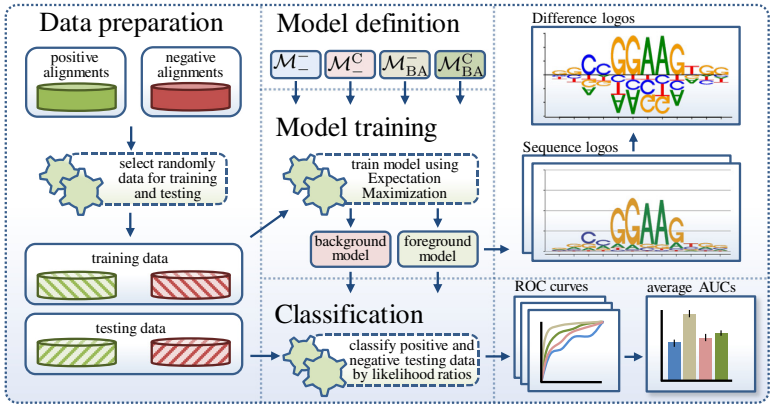


Fig. 5 Overview of the workflow presented in this manuscript. In the data preparation step, we randomly compile disjoint training data and testing data each with positive alignments and negative alignments for each of the transcription factors CTCF, GABP, NRSF, SRF, and STAT1. In the model training step, we train each of the four presented foreground models as well as a background model by expectation maximization with 150 restarts. We choose the foreground model and the background model with maximum likelihood, classify the testing data using a likelihood-ratio classifier, and extract different characteristics such as the ROC curve, the PR curve, the inverse temperature, and the inferred motif. We repeat the described procedure 100 times and calculate mean values and standard errors for several quantities such as the areas under the ROC curves or the PR curves

genome [10, 24], and peak calling has been performed by MACS [32]. Peaks have been extended or cropped to 400 bp, binding regions that potentially comprise more than one of the five transcription factors have been removed, and the 900 binding regions with the highest MACS score have been retained [25]. Orthologous regions from mouse, dog, cow, monkey, horse, and opossum have been extracted from the UCSC database [33], multiple alignments of these orthologous regions have been obtained using T-Coffee [34], and these multiple alignments are kindly provided by [25].

To prepare ungapped alignments from these gapped data sets of the five transcription factors CTCF, GABP, NRSF, SRF, and STAT1, we perform the following three steps. (i) Remove the species that cause the highest number of gaps in all alignments. Accordingly, we remove mouse and opossum and keep orthologous regions from human, monkey, cow, dog, and horse. (ii) Remove all columns in each of the alignments that contain at least one gap to obtain ungapped alignments. (iii) Remove all ungapped alignments that are shorter than 21 bp, which is the length of the longest motif (NRSF) in the performed studies. Table 2 shows details about the resulting data. All data are available as Additional file 2.

Visualizing motif differences with DiffLogo

We used the R package *DiffLogo* [35] to depict the differences between the predicted motifs of the models $\mathcal{M}_{-}^{-}$, $\mathcal{M}_{\mathrm{BA}}^{-}$, $\mathcal{M}_{-}^{\mathrm{C}}$, and $\mathcal{M}_{\mathrm{BA}}^{\mathrm{C}}$. DiffLogo is an open source software that is capable of depicting the differences between multiple motifs [35]. This is realized by visualizing all pairwise differences in an $N \times N$–grid with an empty diagonal. Each entry in the grid is called *difference logo*. The degree of difference of two motifs is calculated by the sum of all stack heights in the corresponding difference logo and is indicated by the background color from red (most dissimilar among all motif pairs) to green (most similar among all motif pairs). The individual sequence logos of the motifs are shown above the table.

A single difference logo depicts the position-specific differences between the base distributions of two sequence motifs. Differences are visualized using a stack of bases for each motif position. The height of each base stack is calculated by the Jensen-Shannon divergence, which is proportional to the degree of base distribution dissimilarity. The Jensen-Shannon divergence is zero if both base distributions are identical, increases with increasing difference of the two base distributions, and reaches a maximum of 2 bit if the two base distributions are maximally different, i.e., if two bases occur only in one of the two motifs each with a probability of 1/2 and the other two bases occur only in the other motif each with a probability of 1/2. The height of each base within a stack is given by the difference of abundance. Thus, the height of a base is proportional to the degree of differential symbol abundance. Bases with a positive height indicate a gain of abundance and bases with a negative height indicate a loss of abundance. The stack height in the positive direction must be equal to the stack height in the negative direction, because the sum of base abundance gain must be equal to the sum of base abundance loss.

Additional files

Additional file 1: Supplementary Methods, Results, Figures, and Examples. This file is structured in four sections.
In section 1, *Modeling the binding-affinity bias*, we describe how to determine the likelihood of non-motif-bearing and motif-bearing alignments modeling the contamination bias and the binding-affinity bias.
In section 2, *Example interpretation of difference logos*, we give an exemplary interpretation of some difference logos.
Section 3, *Supplementary Figures*, contains supplementary Figures S1-S18.
Section 4, *Supplementary Tables*, contains supplementary Tables S1-S10.
(PDF 3492 kb)

Additional file 2: Sequence data. This archive contains data files of gap-free alignments of the ChIP-seq positive regions for each of the transcription factors CTCF, GABP, NRSF, SRF, and STAT1 in FASTA format.
(ZIP 645 kb)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MN and IG developed the key idea. MN and JC developed the computational methods. MN and HT performed the studies. All authors wrote, read, and approved the final manuscript.

Acknowledgements
We thank Lothar Altschmied, Helmut Bäumlein, Sven-Erik Behrens, Karin Breunig, Jan Grau, Katrin Hoffmann, Robert Paxton, Patrice Peterson, and Marcel Quint for valuable discussions and DFG (grant no. GR3526/1), Gencat (2014 SGR 118), and Collectiveware (TIN2015-66863-C2-1-R) for financial support.

Author details
[1]Institute of Computer Science, Martin Luther University, Halle (Saale), Germany. [2]Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany. [3]IIIA-CSIC, Campus UAB, Barcelona, Spain. [4]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany.

Received: 15 December 2015 Accepted: 28 April 2016
Published online: 10 May 2016

References

1. Nowrousian M. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryot Cell. 2010;9(9):1300–10.
2. Villar D, Flicek P, Odom DT. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nat Rev Genet. 2014;15(4):221–33.
3. Park PJ. Chip–seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669–80.
4. Furey TS. Chip–seq and beyond: new and improved methodologies to detect and characterize protein–dna interactions. Nat Rev Genet. 2012;13(12):840–52.
5. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. Chip-seq guidelines and practices of the encode and modencode consortia. Genome Res. 2012;22(9):1813–31.

6. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8. doi:10.1038/nbt.3300.
7. Hawkins J, Grant C, Noble WS, Bailey TL. Assessing phylogenetic motif models for predicting transcription factor binding sites. Bioinformatics. 2009;25(12):339–47.
8. Gomes AL, Abeel T, Peterson M, Azizi E, Lyubetskaya A, Carvalho L, Galagan J. Decoding chip-seq with a double-binding signal refines binding peaks to single-nucleotides and predicts cooperative interaction. Genome Res. 2014;24(10):1686–97.
9. Jain D, Baldi S, Zabel A, Straub T, Becker PB. Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. Nucleic Acids Res. 2015;43(14):6959–68. doi:10.1093/nar/gkv637.
10. Valouev A, Johnson A, David S and. Sundquist, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods. 2008;5(9): 829–34.
11. Rye MB, Sætrom P, Drabløs F. A manually curated chip-seq benchmark demonstrates room for improvement in current peak-finder programs. Nucleic Acids Res. 2011;39(4):e25. doi:10.1093/nar/gkq1187.
12. Jung YL, Luquette LJ, Ho JWK, Ferrari F, Tolstorukov M, Minoda A, Issner R, Epstein CB, Karpen GH, Kuroda MI, Park PJ. Impact of sequencing depth in ChIP-seq experiments. Nucleic Acids Res. 2014;42(9): 178–4. doi:10.1093/nar/gku178.
13. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. Model-based analysis of chip-seq (macs). Genome Biol. 2008;9(9):137.
14. Nix DA, Courdy SJ, Boucher KM. Empirical methods for controlling false positives and estimating confidence in chip-seq peaks. BMC Bioinformatics. 2008;9(1):523.
15. Bailey TL, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol. 2013;9(11):e1003326.
16. Håndstad T, Rye MB, Drabløs F, Sætrom P. A ChIP-Seq Benchmark Shows That Sequence Conservation Mainly Improves Detection of Strong Transcription Factor Binding Sites. PLoS ONE. 2011;6(4):18430. doi:10.1371/journal.pone.0018430.
17. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):51.
18. Park D, Lee Y, Bhupindersingh G, Iyer VR. Widespread Misinterpretable ChIP-seq Bias in Yeast. PloS One. 2013;8(12):83506.
19. Teytelman L, Thurtle DM, Rine J, van Oudenaarden A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc Nat Acad Sci. 2013;110(46):18602–7.
20. Elliott JH, Grimshaw J, Altman R, Bero L, Goodman SN, Henry D, Macleod M, Tovey D, Tugwell P, White H, Sim I. Informatics: Make sense of health data. Nature. 2015;527:31–2.
21. Bailey TL, Elkan C. The value of prior knowledge in discovering motifs with MEME. Ismb. 1995;3:21–9.
22. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in chip-seq peak detection. PloS one. 2010;5(7):11471.
23. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. J Mol Biol. 1986;188(3):415–31.
24. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucl Acids Res. 2008;36(16):5221–31. doi:10.1093/nar/gkn488. http://nar.oxfordjournals.org/cgi/reprint/36/16/5221.pdf.
25. Arnold P, Erb I, Pachkov M, Molina N, van Nimwegen E. MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. Bioinformatics. 2012;28(4):487–94. doi:10.1093/bioinformatics/btr695.
26. Sober E. The principle of parsimony. Brit J Philos Sci. 1981;32(2):145–56.
27. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen C-YY, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A, Wasserman WW. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2014;42(Database issue):142–7. doi:10.1093/nar/gkt997.
28. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368–76.
29. Lawrence CE, Reilly AA. An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins: Struct Funct Bioinformatics. 1990;7(1): 41–51.
30. Quantitative Enrichment of Sequence Tags: QuEST. http://mendel.stanford.edu/sidowlab/downloads/quest/. Accessed 29 Mar 2016.
31. ChIP-Seq Data Analysis: Identification of Protein–DNA Binding Sites with SISSRs Peak-Finder. http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/sissrs/. Accessed 29 Mar 2016.
32. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-seq (MACS). Genome Biol. 2008;9(9):137. doi:10.1186/gb-2008-9-9-r137.
33. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Kober KM, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ. The UCSC genome browser database: 2008 update. Nucleic Acids Res. 2008;36(suppl 1):773–9. doi:10.1093/nar/gkm966.
34. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000;302(1): 205–17. doi:10.1006/jmbi.2000.4042.
35. Nettling M, Treutler H, Grau J, Keilwagen J, Posch S, Grosse I. DiffLogo: a comparative visualization of sequence motifs. BMC Bioinf. 2015;16(1):1.

5.2 Unrealistic phylogenetic trees may improve phylogenetic footprinting

M Nettling, H Treutler, J Cerquides, I Grosse. 2017. Unrealistic phylogenetic trees may improve phylogenetic footprinting. *Bioinformatics*
doi:10.1093/bioinformatics/btx033

Nettling *et al. BMC Bioinformatics* (2017) 18:141
DOI 10.1186/s12859-017-1495-1

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Combining phylogenetic footprinting with motif models incorporating intra-motif dependencies

Martin Nettling[1*], Hendrik Treutler[2], Jesus Cerquides[3] and Ivo Grosse[1,4]

Abstract

Background: Transcriptional gene regulation is a fundamental process in nature, and the experimental and computational investigation of DNA binding motifs and their binding sites is a prerequisite for elucidating this process. Approaches for de-novo motif discovery can be subdivided in phylogenetic footprinting that takes into account phylogenetic dependencies in aligned sequences of more than one species and non-phylogenetic approaches based on sequences from only one species that typically take into account intra-motif dependencies. It has been shown that modeling (i) phylogenetic dependencies as well as (ii) intra-motif dependencies separately improves de-novo motif discovery, but there is no approach capable of modeling both (i) and (ii) simultaneously.

Results: Here, we present an approach for de-novo motif discovery that combines phylogenetic footprinting with motif models capable of taking into account intra-motif dependencies. We study the degree of intra-motif dependencies inferred by this approach from ChIP-seq data of 35 transcription factors. We find that significant intra-motif dependencies of orders 1 and 2 are present in all 35 datasets and that intra-motif dependencies of order 2 are typically stronger than those of order 1. We also find that the presented approach improves the classification performance of phylogenetic footprinting in all 35 datasets and that incorporating intra-motif dependencies of order 2 yields a higher classification performance than incorporating such dependencies of only order 1.

Conclusion: Combining phylogenetic footprinting with motif models incorporating intra-motif dependencies leads to an improved performance in the classification of transcription factor binding sites. This may advance our understanding of transcriptional gene regulation and its evolution.

Keywords: ChIP-Seq, Phylogenetic footprinting, Evolution, Transcription factor binding sites, Gene regulation

Background

Gene regulation is an essential process in every living organism that controls the activity of gene expression and enables the concerted up- and down-regulation of gene products. Gene regulation involves a wide range of sub-processes such as transcriptional regulation including DNA methylation [1], histon modifications [2], and promotor escaping [3] as well as post-transcriptional regulation including modulated mRNA decay [4], siRNA interference [5, 6], and alternative splicing [7, 8]. One important process in gene regulation is the interaction of transcription factors (TFs) with their corresponding transcription factor binding sites (TFBSs) [9, 10]. The algorithmic discovery of TFBSs and the simultaneous inference of their motifs is known as de-novo motif discovery and a challenging task in bioinformatics. Many different approaches exist for de-novo motif discovery, which can be divided in two groups.

The first group comprises approaches based on sequences of only one species, which we refer to as one-species approaches in this work, using statistical models for the binding of TFs to their TFBSs. One of the most popular motif models is the simple position weight matrix (PWM) model, which does not take into account any dependency between different positions of the same TFBS, but there are also more complex motif models that

*Correspondence: martin.nettling@informatik.uni-halle.de
[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle, Germany
Full list of author information is available at the end of the article

take into account intra-motif dependencies. Irrespective of the wide variety of different motif models used, all of these approaches have in common that they do not take into account phylogenetic information available from orthologous sequences of phylogenetically related species.

Complex motif models that take into account intra-motif dependencies have been shown to outperform simpler motif models like the PWM model [11–13]. Examples for highly popular tools that model intra-motif dependencies are *Dimont* [14], *MEME-ChIP* [15], *DeepBind* [16], and *diChIPMunk* [17].

In contrast, the second group of de-novo motif discovery approaches known as phylogenetic footprinting incorporates orthologous sequences of at least two phylogenetically related species. The basic idea of these approaches is that TFBSs tend to be subject to negative selection during evolution, which can increase the recognition of TFBSs in the reference species. Phylogenetic motif models, which model the binding of TFs to their TFBSs and their evolution simultaneously, are based on evolutionary models such as the popular Felsenstein model [18]. Irrespective of the wide variety of different phylogenetic motif models used, all of these approaches have in common that they do not take into account intra-motif dependencies.

Not all sequences from the reference species may have orthologous sequences in phylogenetically related species, and not all aligned sequences may comprise functional TFBSs at the same alignment positions [19]. Moreover, alignment errors, binding site turnovers, and spurious alignments from convergent evolution may affect the utility of phylogenetic footprinting. Nevertheless, phylogenetic footprinting has been shown to outperform one-species approaches for many TFs and have become increasingly attractive due to next generation sequencing and the resulting avalanche of data [20–22].

Examples for highly popular phylogenetic footprinting tools that have been applied to eukaryotes and prokaryotes are *FootPrinter* [23], *PhyME* [24], *MONKEY* [25], *MicroFootprinter* [26], *Phylogenetic Gibbs Sampler* [27], *PhyloGibbs* [28], *PhyloGibbs-MP* [29], or *MotEvo* [30].

In summary, one-species approaches neglect phylogenetic information, whereas phylogenetic footprinting, which incorporates this information, neglects intra-motif dependencies. The main objective of this work is to develop an approach that combines these two ideas and to investigate if taking into account intra-motif dependencies can improve phylogenetic footprinting. Specifically, we propose a simple phylogenetic footprinting model (PFM) capable of taking into account both intra-motif dependencies and phylogenetic information in Methods, and we study if modeling intra-motif dependencies improves phylogenetic footprinting based on human ChIP-Seq data of 35 TFs and more than 10^5 multiple alignments of human ChIP-seq positive regions and their orthologous sequences of 9 mammalian species ranging from chimp to cow in Results.

Methods

In this section we describe (i) the studied datasets, (ii) the used notation and the likelihood calculation of the PFM, (iii) the performance measure, (iv) the calculation of the mutual information, and (v) details regarding the estimation algorithm and implementation of the proposed model.

Data

We use freely available ChIP-Seq data for 50 transcription factors from the ENCODE project [31, 32]. The ChIP-seq experiments were performed by several production groups in the ENCODE Consortium and analysed by the ENCODE Analysis Working Group based on a uniform processing pipeline developed for the ENCODE Integrative Analysis effort [33]. We focus on datasets for the human H1-hESC cell line. The uniform processing pipeline utilizes the SPP peak caller [34] and biological replicates (at least two per transcription factor) are analysed jointly with a Irreproducible Discovery Rate (IDR) score of at least 2%. The resulting ChIP-seq regions of the Uniform TFBS track reference the hg19 assembly [35] and each comprise the chromosome, start position, end position, and an enrichment score. We exclude 15 datasets which yield repetitive motifs analog to [13] and hence retain datasets of 35 TFs.

For each TFs we select the top 20% of the available ChIP-seq regions ranked by enrichment score. We denote these regions as ChIP-seq positive regions and use them as basis for the positive dataset (Additional file 1: Table S1 and Additional file 1: Section 1.3). We denote the regions between ChIP-seq positive regions on one chromosome as ChIP-seq negative regions. For each TF we extract two regions of length 500 bp from each ChIP-seq negative region centered at one third and two thirds, and use these as basis for the negative dataset. Hence, there are roughly twice as many negative regions than positive regions. We remove regions from the positive and the negative region sets that are shorter than 20 bp. For each region in the positive and negative region sets we extract the corresponding alignment consisting of 46 mammals using the freely available multiple genome alignment from UCSC [36].

We apply the following steps to each alignment. We remove alignment columns with gap-symbols or ambiguous symbols in the human sequence and concatenate the remaining alignment columns. We retain the 10 species with the best alignment coverage, namely Human (hg19), Chimp (panTro), Baboon (papHam), Orangutan (ponAbe), Rhesus (rheMac), Marmoset (calJac), Horse, (equCab), Dog (canFam), Gorilla (gorGor), and Cow (bosTau).

We replace ambiguous symbols with gap-symbols. We remove all alignments which comprise no base symbols for 20% or more species. See Additional file 1: Table S1 for statistics on the number of ChIP-Seq positive regions and the number of extracted alignments and see Additional file 1: Table S2 for details about the origin of the used ChIP-Seq data and Additional file 2 contains all extracted alignments.

Phylogenetic footprinting model

Notation

Each dataset of each TF contains N alignments, with each alignment containing O sequences (one per observed species). Of course the number of alignments per TF, N, varies from TF to TF (See Additional file 1: Table S1). The n-th alignment is denoted by X_n and its length is denoted by L_n. Each sequence of alignment X_n is composed of L_n symbols. We denote by $X_n^{u,o}$ the u-th symbol of the o-th sequence of the n-th alignment. All symbols belong to the set $\mathcal{A} = \{A, C, G, T, -\}$ where A, C, G, and T denote the bases and $-$ denotes a gap in the alignment. Missing sequences in alignment n are represented by L_n gap symbols.

An alignment X_n may or may not contain a binding site. This is encoded in the variable M_n, with $M_n = 0$ indicating that alignment X_n does not contain a motif and $M_n = 1$ indicating that alignment X_n does contain a motif. This model is known as *ZOOPS* (zero or one occurrence of a binding site per sequence) or *NOOPS* (noisy OOPS) model. Due to its simplicity and its modularity this model is widely used for de-novo motif discovery [37–40].

Likelihood

The probability that the alignment X_n is generated by our PFM can be written as

$$\begin{aligned} p\left(X_n|\theta\right) = & p\left(X_n|M_n = 0, \theta\right) \cdot p\left(M_n = 0|\theta\right) \\ & + p\left(X_n|M_n = 1, \theta\right) \cdot p\left(M_n = 1|\theta\right) \end{aligned} \tag{1}$$

with variable M_n taking a Bernoulli distribution and θ denoting model parameters, namely (i) the topology of the phylogenetic tree, (ii) the substitution probabilities, and (iii) the evolutionary model with its stationary probabilities for the flanking regions as well as for the binding site regions.

We need to specify the probability for non-motif-bearing $p(X_n|M_n = 0, \theta)$ and for motif-bearing alignments $p(X_n|M_n = 1, \theta)$. For reasons of clarity we omit θ in the following.

Likelihood of a non-motif-bearing alignment

Since sequences are assumed to be conditionally independent, the probability of an alignment decomposes as the product of the probability of each of its sequences:

$$p\left(X_n|M_n = 0\right) = \prod_{o=1}^{O} p\left(X_n^{:,o}|M_n = 0\right) \tag{2}$$

Now, the probability of each sequence follows a homogeneous Markov Chain of order C:

$$p\left(X_n^{:,o}|M_n = 0\right) = \prod_{u=1}^{L_n} p\left(X_n^{u,o}|X_n^{p(u,1),o}, M_n = 0\right), \tag{3}$$

where $p(u, k)$ stands for the (at most C) predecessors of the u-th base for a sequence starting at position k, namely the set $p(u, k) = \{v| \max(k, u - C) \leq v < u\}$, and

$$p\left(X_n^{u,o} = a|X_n^{p(u,1),o} = \zeta, M_n = 0\right) = \pi_0^{a,\zeta} \tag{4}$$

where $\pi_0^{a,\zeta}$ is the parameter encoding the probability of a base a in the background sequence provided that its predecessors are in joint state ζ.

Likelihood of a motif-bearing alignment

We note W for the length of the motif. Since the motif can be present in different positions, the probability of a motif-bearing assignment is a weighted sum over each possible motif position ℓ_n:

$$\begin{aligned} p\left(X_n|M_n = 1\right) = & \sum_{\ell_n=1}^{L_n-W+1} p\left(X_n|\ell_n, M_n = 1, \theta, \right) \\ & \times p\left(\ell_n|M_n = 1\right) \end{aligned} \tag{5}$$

We assume motifs to be uniformly distributed a priori, thus having that $p(\ell_n|M_n = 1) = \frac{1}{L_n-W+1}$. Again, conditional independence of sequences allows to express probability of an alignment as a product of the probability of its single sequences

$$p\left(X_n|\ell_n, M_n = 1\right) = \prod_{o=1}^{O} p\left(X_n^{:,o}, \ell_n, M_n = 1\right) \tag{6}$$

And the probability of each single sequence breaks into three parts: (i) an initial non-motif bearing part containing bases $i(\ell_n) = \{1, \ldots, \ell_n - 1\}$, (ii) the motif, containing bases $m(\ell_n) = \{\ell_n, \ldots, \ell_n + W - 1\}$ and (iii) a final non-motif bearing part formed by bases $e(\ell_n) = \{\ell_n + W, \ldots, L_n\}$:

$$\begin{aligned} p\left(X_n^{:,o}|\ell_n, M_n = 1\right) = & p\left(X_n^{i(\ell_n),o}|\ell_n, M_n = 1\right) \\ & \times \left(X_n^{m(\ell_n),o}|\ell_n, M_n = 1\right) \\ & \times p\left(X_n^{e(\ell_n),o}|\ell_n, M_n = 1\right) \end{aligned} \tag{7}$$

with the non-motif bearing parts following a homogeneous Markov Chain of order C as described above

and the motif-bearing part following a non-homogeneous Markov Chain defined as

$$p\left(X_n^{m(\ell_n),o}|\ell_n, M_n = 1\right) = \prod_{u\in m(\ell_n)} p\left(X_n^{u,o}|X_n^{p(u,\ell_n),o}, \ell_n, M_n = 1\right), \quad (8)$$

with

$$p\left(X_n^{u,o} = a|X_n^{p(u,\ell_n),o} = \zeta, \ell_n, M_n = 0\right) = \pi_{u-\ell_n+1}^{a,\zeta} \quad (9)$$

where $\pi_w^{a,\zeta}$ is a parameter that encodes the probability of a base a, at position w of the motif provided that its predecessors are in joint state ζ.

Management of gaps

A sequence may have gaps introduced by the alignment algorithm. We compute the probability of a gap by summing over all possible nucleotides at that position in that sequence. For example to assess $p\left(X_n^{u,o} = -|X_n^{p(u,1),o} = \zeta, M_n = 0\right)$, we use $\sum_{a\in\{A,C,G,T\}} p\left(X_n^{u,o} = a|X_n^{p(u,1),o} = \zeta, M_n = 0\right)$.

The used model estimation procedure and the freely available implementation are specified in Methods 5, and run times are exemplified in Additional file 1: Section 1.6.

Measuring classification performance

We evaluate all PFMs by a stratified repeated random subsampling validation by estimating all PFMs from a training set and measuring classification performance on a test set as follows.

In step 1, we generate two training sets and two disjoint test sets for each of the 35 transcription factors as follows. We randomly select 70% but maximal 1000 alignments from the set of alignments of a particular transcription factor as positive training set, and we choose the set of the remaining alignments but maximal 1000 as positive test set. We randomly select 70% but maximal 1000 alignments from the corresponding set of negative alignments of this transcription factor, and we choose the set of the remaining alignments but maximal 1000 as negative test set.

In step 2, we train a foreground model on the positive training set and a background model on the negative training set by expectation maximization [41] using a numerical optimization procedure in the maximization step. In all cases, we attempt to find a motif of length $W = 20$ bp. It is known that the motifs of many TFs have a length smaller than W bp, but adding some possibly uninformative positions in case of short motifs is less harmful than not being able to take into account all motif positions in case of long motifs. We restart the expectation maximization algorithm, which is deterministic for a given dataset and a given initialization, 100 times with different initializations and choose the foreground model and the background model with the maximum likelihood on the positive training data and the negative training data, respectively, for classification. We use a likelihood-ratio classifier of the two chosen foreground and background models, apply this classifier to the disjoint positive and negative test sets, and calculate the area under the receiver operating characteristics curve and the area under the precision recall curve as measures of classification performance.

We repeat both steps 25 times and determine (i) the mean area under the receiver operating characteristic curve and its standard error and (ii) the mean area under the precision recall curve and its standard error.

Relative increase of classification performance

We compute the relative increase or decrease of the classification performance of the PFM(1) and the PFM(2) relative to the PFM(0), where PFM(C) denotes a PFMs taking into account base dependencies of order C. We compute $R_{PFM(C)}$ as the ratio of the improvement of the PFM(C) relative to the PFM(0) divided by the maximum possible improvement to the PFM(0) as given by

$$R_{PFM(C)} = \frac{AUC_{PFM(C)} - AUC_{PFM(0)}}{1 - AUC_{PFM(0)}}.$$

Negative values of $R_{PFM(C)}$ denote a decrease of classification performance and positive values of $R_{PFM(C)}$ denote an increase of classification performance up to a maximum of $R_{PFM(C)} = 1$ which denotes perfect classification (provided that the AUC of PFM(0) is smaller than 1).

Mutual information

The mutual information (MI) is a standard measure for quantifying statistical dependencies. We compute the MI between a base at position w in a motif and its C preceding bases for $w > C$ as follows

$$I_C(w) = I\left(X_w, X_w^C\right) = \sum_{a\in\mathcal{A}^C}\sum_{b\in\mathcal{A}} p\left(X_w^C = a, X_w = b\right) \times \log_2 \frac{p\left(X_w^C = a, X_w = b\right)}{p\left(X_w^C = a\right)p(X_w = b)}$$

where X_w denotes the base at position w and $X_w^C = (X_{w-C}, \ldots, X_{w-1})$ denotes the context of X_w. $I_C(w)$ denotes the amount of information in the C-mer ending at position $w-1$ about its adjacent base at position w. $I_C(w)$ is undefined for $w \leq C$.

We denote the vector of MIs values $I_C(w)$ for $w \in \{C+1, \ldots, W\}$ by $I_C = (I_C(C+1), \ldots, I_C(W))$, where W is the length of the motif, and we call this vector MI profile.

Implementation

We implement the proposed PFM based on the freely available Java Framework *Jstacs* [42]. Among others, *Jstacs* provides ready-to-use sequence models for reuse, numerical and non-numerical optimization procedures for model estimation, serialization of models, and methods for the statistical evaluation of results. In contrast to existing tools which are typically focused on application, using *Jstacs* we are able to compare different PFMs in a detailed way by extracting mandatory information about the inferred models and the predicted binding sites.

Algorithm 1 shows the pseudocode for inferring a PFM from a set of alignments. The implementation of the proposed phylogenetic footprinting model is available at https://github.com/mgledi/PhyFoo/.

Algorithm 1 Motif discovery algorithm for the proposed PFM. Upon random initialization of the model parameters we iteratively estimate sequence weights and model parameters with multiple algorithm restarts, where R denotes the number of restarts of the whole algorithm, and S denotes the number of iterations. The result is the set of model parameters with maximum likelihood

Data: Set of alignments $\{X_1, \ldots, X_N\}$
for r = 1 ... R **do**
  Initialize θ^1 randomly
  for s = 1 ... S **do**
    E-step: Estimate $p(X_n^{m(\ell_n),o}|\ell_n, M_n = 1, \theta^s)$ for each position ℓ_n in each alignment X_n given the model parameters θ^s (see Eq. 8)
    M-step: Maximize $p(X_n|\theta^{s+1})$ regarding θ^{s+1} given all alignments and the probabilities $p(X_n^{m(\ell_n),o}|\ell_n, M_n = 1, \theta^s)$ (see Eq. 1)
  end for
  Keep θ^{S+1} denoted θ_r
end for
Result: $\theta \in \{\theta_1, \ldots \theta_R\}$ with maximum likelihood

Results and discussion

We propose a model for phylogenetic footprinting that is capable of taking into account intra-motif dependencies as specified in Methods 2. Specifically, we model intra-motif dependencies in TFBSs as well as dependencies among adjacent bases in flanking sequences by Markov models of orders 0, 1, and 2, and we denote the proposed PFM by PFM(0), PFM(1), and PFM(2).

In the first subsection we study if the proposed PFMs can capture intra-motif dependencies of orders 1 and 2 in ChIP-Seq data of 35 TFs. In the second subsection we study if modeling base dependencies can improve phylogenetic footprinting. Both studies are based on human sequences extracted from ENCODE ChIP-seq data [33] and corresponding orthologous sequences of 9 mammalian species, yielding 35 data sets comprising 135196 multiple sequence alignments with an average length of 124 bases (Methods 1).

Intra-motif dependencies can be captured by phylogenetic footprinting

In this subsection we study to which degree intra-motif dependencies can be captured using the PFMs of orders 1 and 2.

We measure the degree of intra-motif dependencies of order 1 between two neighboring bases or of order 2 between a dimer and its neighboring base by the MI as described in Methods 4. The MI quantifies the amount of information in a base or a dimer about the neighboring base in units of bits and ranges from 0 bits in case of statistical independence to 2 bits in case of deterministic dependency of the considered base on the preceding base or the preceding dimer. We compute the MI for every position of a binding site and call the resulting vector of MI values MI profile.

For each of the 35 TFs, we compute the two MI profiles of orders 1 and 2 from the motifs obtained by phylogenetic footprinting using the PFM(2). We present the resulting 35×2 MI profiles as Additional file 3 and the 2×2 MI profiles of the two TFs *CJUN* and *Nrf* as examples in Fig. 1a.

First, we study the MI profiles of order 1 for these two TFs. For both TFs we find statistically significant intra-motif dependencies between neighboring bases at all positions. For *CJUN*, intra-motif dependencies of order 1 are particularly strong at motif positions 2 to 4, yielding a maximum MI of 0.52 bits at motif position 4. For *Nrf*, intra-motif dependencies of order 1 are particularly strong at motif positions 8 to 11 and 14 to 15, yielding a maximum MI of 0.23 bits at motif position 11.

Next, we study the MI profiles of order 2. Again, we find statistically significant intra-motif dependencies between dimers and their neighboring bases at all positions for both *CJUN* and *Nrf*. For *CJUN*, intra-motif dependencies of order 2 are particularly strong at motif positions 2 to 4, yielding a maximum MI of 0.70 bits at motif position 3. For *Nrf*, intra-motif dependencies of order 2 are particularly strong at motif positions 8 to 11 and 13 to 15, yielding a maximum MI of 0.28 bit at motif position 11.

Moreover, we find that intra-motif dependencies of order 2 are significantly stronger than the corresponding intra-motif dependencies of order 1 at several positions for both *CJUN* and *Nrf*. Comparing the MI profiles of orders 1 and 2, we find that the MI profile of order 2 is up to twofold higher than the MI profile of order 1 for *CJUN* and up to sevenfold higher for *Nrf*, stating that in both TFs there are significant intra-motif dependencies of

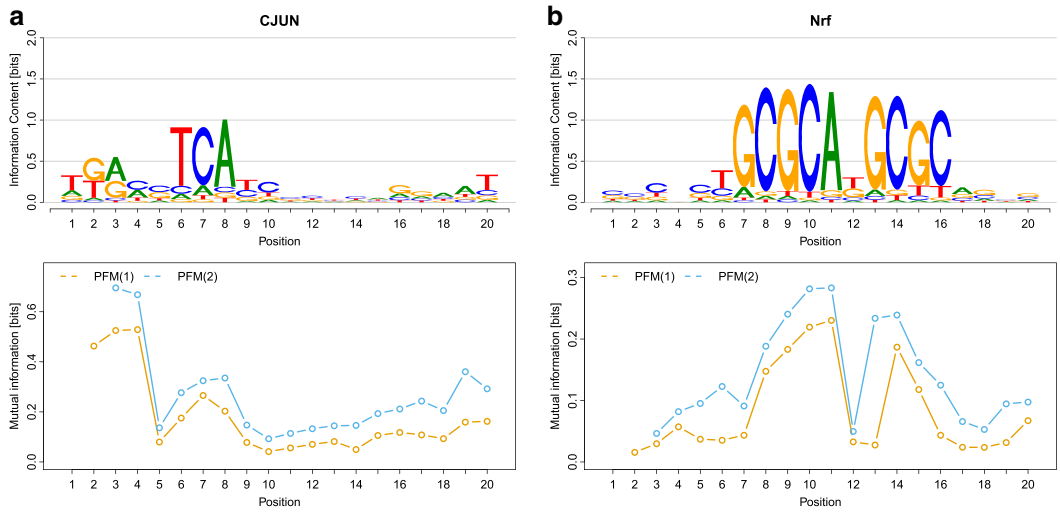


Fig. 1 Sequence logos and intra-motif dependencies for the TFs **a** *CJUN* and **b** *Nrf*. We depict for both TFs (i) the sequence logo inferred by the PFM(2) from all species in the first row and (ii) the MI profiles of orders 1 and 2 inferred by the PFM(2) in the second row. The MI profiles of order 2 are larger than the MI profiles of order 1. Please see Additional file 3 for the MI profiles of all 35 TFs and Additional file 5 for all sequence logos of all 35 TFs for the PFMs of orders 0, 1, and 2

order 2 beyond those expected from the corresponding intra-motif dependencies of order 1.

Next, we study the MI profiles of orders 1 and 2 for all 35 TFs. In order to condense the results and to allow a visual comparison of the results for both profiles and all 35 TFs, we show for each MI profile and each TF the maximum and mean MI values in Fig. 2a.

We find that the average of the 35 maximum MI values of order 1 is 0.39 bits, whereas the average of the 35 maximum MI values of order 2 is significantly greater at 0.56 bits. Likewise, we find that the average of the 35 mean MI values of order 1 is 0.14 bits, whereas the average of the 35 mean MI values of order 2 is significantly greater at 0.23 bits. These observations suggest that intra-motif dependencies are present in all of the studied TFs and that intra-motif dependencies of order 2 are typically stronger than those of order 1.

By scrutinizing Figs. 2a and b, however, we also find that the maximum and mean MIs values vary significantly from TF to TF. For example, we find a maximum and mean MI value of order 1 of 0.11 bits and 0.05 bits for *CEBPB* and a maximum and mean MI value of order 1 of 0.89 bits and 0.20 bits for *Mxi*. Analogously, we find a maximum and mean MI value of order 2 of 0.16 bits and 0.07 bits for *CEBPB* and a maximum and mean MI value of order 2 of 1.15 bits and 0.37 bits for *Mxi*.

To study the possibility that these captured intra-motif dependencies are an artifact resulting from a mixture of different species-specific motifs, we finally study the similarity of the 10 species-specific motifs as well as the 20 species-specific MI profiles of orders 1 and 2. We find that the observed pairwise differences between the species-specific motifs are not significant (Additional file 1: Section 1.1.1). Moreover, we find that the species-specific MI profiles are similar to each other and to the corresponding MI profiles captured by phylogenetic footprinting (Additional file 4, Additional file 1: Section 1.1.2). Both findings indicate that the intra-motif dependencies shown in Fig. 1b and in Additional file 3 cannot be explained as an artifact resulting from a mixture of different species-specific motifs.

Modeling intra-motif dependencies improves phylogenetic footprinting

In this subsection we study if modeling base dependencies can improve phylogenetic footprinting.

First, we compute the classification performance of the PFMs of orders 0, 1, and 2 as described in Methods 3. Second, we determine the increase of the classification performance of the PFMs taking into account base dependencies of orders 1 and 2 relative to the classification performance of the PFM neglecting base dependencies as described in Methods 3. Here, positive values indicate an increase of classification performance, while negative values indicate a decrease of classification performance.

Figure 3a shows the classification performances of the PFMs of orders 0, 1, and 2 for each of the 35 TFs, and Fig. 3b shows the corresponding relative increases. We

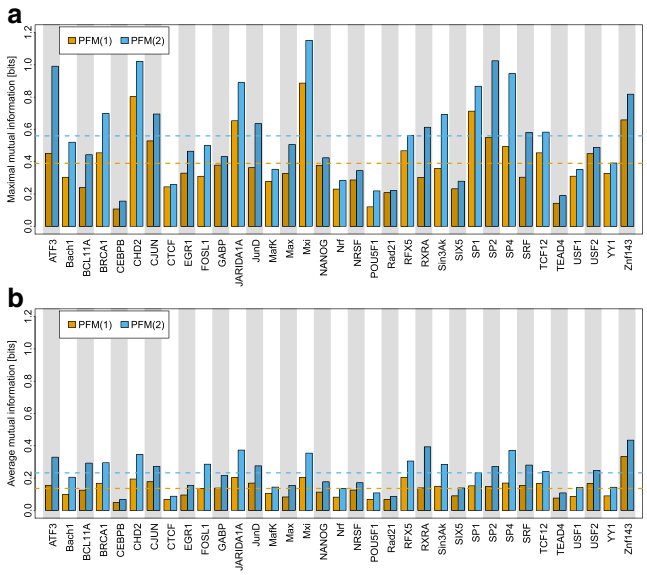


Fig. 2 Maximum and average MIs of MI profiles inferred by the PFM(2) for all 35 TFs. In Fig. **a** we show the maximum MI of the MI profiles of orders 1 and 2. In Fig. **b** we show the average MI of the MI profiles of orders 1 and 2. The dashed lines indicate the mean of the maximum MIs and the mean of the average MIs for both MI profiles respectively. The degree of intra–motif dependencies depends of the TF and is always larger in case of intra–motif dependencies of order 2. Please see Additional file 3 for the MI profiles of all 35 TFs

find that modeling base dependencies of order 1 increases the classification performance in 31 of 35 cases, and we find that modeling base dependencies of order 2 increases the classification performance in all of the 35 cases. Moreover, we find that modeling base dependencies of order 2 always yields a higher classification performance than modeling base dependencies of order 1.

By scrutinizing Fig. 3a, we find that the differences of the classification performances of the PFMs of orders 1 and 2 and the PFMs of order 0 vary significantly from TF to TF. For example, in case of base dependencies of order 1 we find the highest difference of 11% for CHD2 and the lowest difference of −1% for Rad21. In case of base dependencies of order 2 we find the highest difference of 13% for Rad21 and the lowest difference of 1% for RXRA.

By scrutinizing Fig. 3b, we find that also the relative increases of classification performances vary significantly from TF to TF. For example, in case of base dependencies of order 1 we find the highest increase of 70% for JARIDA1A and the lowest increase of −7% for Rad21. In case of base dependencies of order 2 we find the highest increase of 78% for JARIDA1A and the lowest increase of 7% for RXRA.

Figure 4 summarizes the results by showing (a) the classification performance of the PFMs of orders 0, 1, and 2 averaged over all 35 TFs and (b) the relative increases of classification performances averaged over all 35 TFs. We observe that the average classification performance increases significantly from order 0 to order 1 and from order 1 to order 2. Specifically, we find that the average classification performance of the PFM(1) is 4.6% higher than that of the PFM(0) and that the average classification performance of the PFM(2) is 3.5% higher than that of the PFM(1). We find that the average relative increase of the classification performance of the PFM(1) over that of the PFM(0) is 25% and that the average relative increase of the classification performance of the PFM(2) over that of the PFM(0) is 42%.

Next, we study the robustness of the proposed approach with respect to the number of species in the multiple sequence alignments. We perform the same study on the same 35 datasets with alignments comprising only subsets of the 10 species, and we find that for all subsets the classification performance increases significantly from order 0 to order 1 for many of the 35 TFs and from order 1 to order 2 for all of the 35 TFs (Additional file 1: Section 1.2).

These findings indicate that taking into account base dependencies improves phylogenetic footprinting, but they also indicate that this improvement is small. Given the fact that taking into account base dependencies improves one-species approaches, too, it could well be that the improvement obtained by taking into account

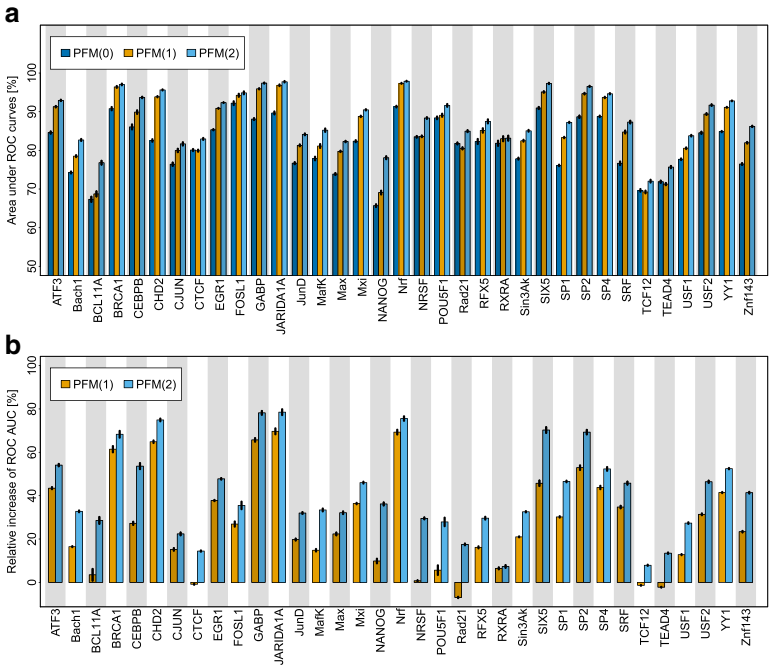


Fig. 3 Classification performance for PFMs with base dependencies of orders 0, 1 and 2. **a** We show the mean and standard error of the ROC AUC for PFMs of orders 0, 1, and 2 averaged over 25–fold stratified repeated random subsampling. **b** We plot the mean and standard error of the relative increase of the ROC AUC for the PFMs of orders 1 and 2 relative to the PFM or order 0 for each of the 35 TFs. Taking into account base dependencies of order 1 increases the classification performance for 31 TFs. Taking into account base dependencies of order 2 increases the classification performance in all cases and is larger compared to taking into account base dependencies of order 1 in all cases. See Additional file 6 for detailed ROC and PR curves for the PFMs of order 2

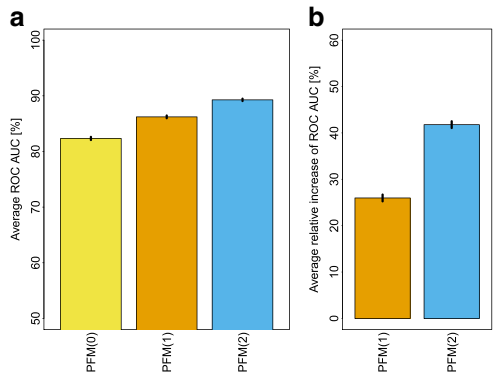


Fig. 4 Classification performance averaged for all 35 TFs. **a** We show the ROC AUC for PFMs of orders 0, 1, and 2 in percent averaged over 25–fold stratified repeated random subsampling and averaged over all 35 TFs. The overall classification performance increases with the order of the PFM. **b** We show the improvement of the ROC AUC for the PFMs of orders 1 and 2 relative to the PFM of order 0 averaged over 25–fold stratified repeated random subsampling and averaged over all 35 TFs

base dependencies in one-species approaches is greater than in phylogenetic footprinting. Such a difference could result in the situation where the advantage of phylogenetic footprinting over one-species approaches when neglecting base dependencies decreases or even turns into a disadvantage when taking into account base dependencies.

To study to which degree the small improvement of phylogenetic footprinting by taking into account base dependencies might be overshadowed by a possibly greater improvement of one-species approaches, we compare the classification performances of the four cases of one-species approaches and phylogenetic footprinting when neglecting and taking into account base dependencies (Additional file 1: Section 1.3). Consistent to previous studies, we find that phylogenetic footprinting yields a higher (lower) classification performance compared to one-species approaches for 23 (12) of the 35 TFs when neglecting base dependencies. When taking into account base dependencies, however, phylogenetic footprinting yields a higher (lower) classification performance compared to one-species approaches in 31 (4) of the 35 TFs.

This finding indicates that the small improvement of phylogenetic footprinting by taking into account base dependencies is greater than the corresponding improvement of one-species approaches. It also indicates that the previously observed advantage of phylogenetic footprinting over one-species approaches when neglecting base dependencies (23 to 12) does not decrease or turn into a disadvantage, but becomes even more pronounced (31 to 4), when taking into account base dependencies. This increased advantage of phylogenetic footprinting over one-species approaches achieved by taking into account base dependencies is surprising as it indicates the presence of some synergy of modeling both phylogenetic and base dependencies.

We finally study for each of the 35 TFs which of the four models yields the highest classification performance, and we find that one-species approaches neglecting base dependencies yields the highest classification performance for one TF (*CEBPB*), one-species approaches taking into account base dependencies yields the highest classification performance for three TFs (*BCL11A*, *MafK*, and *RXRA*), phylogenetic footprinting neglecting base dependencies never yields the highest classification performance, and phylogenetic footprinting taking into account base dependencies yields the highest classification performance for 31 TFs. This finding indicates that phylogenetic footprinting can be improved by taking into account base dependencies, that one-species approaches using base dependencies can be improved by taking into account phylogenetic dependencies, and that there is a surprising synergy of simultaneously modeling both phylogenetic and base dependencies.

Conclusions

In this work, we introduced a phylogenetic footprinting model capable of taking into account base dependencies and evaluated this phylogenetic footprinting model on ChIP-seq data of 35 TFs. We found significant intra-motif dependencies of orders 1 and 2 in all 35 datasets and that the inferred intra-motif dependencies of order 2 are stronger than those of order 1 for all 35 TFs. We also found that these intra-motif dependencies cannot be explained as an artifact resulting from a mixture of different species-specific motifs. We further found that the classification performance of the introduced phylogenetic footprinting model is higher than that of phylogenetic footprinting models neglecting base dependencies for all of the 35 TFs and higher than that of one-species approaches for 31 of the 35 TFs. These findings suggest that combining phylogenetic footprinting with motif models incorporating intra-motif dependencies may lead to an improved prediction of TFBSs and thus advance our understanding of transcriptional gene regulation and its evolution.

Additional files

Additional file 1: Supplementary Material. This file is structured in three sections, presenting four additional studies, details about the implementation and some statistics regarding the datasets of all 35 TFs. In Section 1, *Supplementary Results*, we first study differences among species–specific motifs of 35 TFs. We then study the robustness of the proposed PFM to different species compositions on data of 35 TFs. Third, we examine the impact of base dependencies and phylogenetic dependencies on classification performance. In the fourth subsection, we compare the proposed PFM(2) with a state of the art tool by Eggeling et al. 2015 [13] on data of 35 TFs. In the fifth subsection, we show statistics of the distances between ChIP-seq positive regions and the alignment coverage of ten species. Finally, we specify the run–time of our freely available implementation of the proposed PFM.
In Section 2, *Supplementary Methods*, we specify details about the estimation of species–specific motifs and we define a statistical test for the significance of differences among species–specific motifs.
In Section 3, *Supplementary Tables*, we show statistics of the datasets of 35 TFs, summarize results regarding the significance of species–specific motifs and the impact of base dependencies and phylogenetic dependencies, and show the alignment coverage of ten species for 35 TFs. (PDF 1034.24 kb)

Additional file 2: Sequence data. This archive contains data files of alignments of the ChIP-seq positive regions and negative control regions for each of the 35 TFs in FASTA format. (ZIP 83763.2 kb)

Additional file 3: Sequence logos, MI profiles of order 1, MI profiles of order 2, and species-specific MI profiles of orders 1 and 2. The file contains for each of the 35 TFs the sequence logo inferred using the PFM(2) aligned with MI profiles of order 1, the MI profiles of order 2, and species-specific MI profiles of orders 1 and 2 for each of the 10 species. (PDF 2129.92 kb)

Additional file 4: Tables of difference logos. The file contains for each of the 35 TFs a 10 × 10 table of difference logos for a pair-wise visual comparison of species-specific motifs. (ZIP 26112 kb)

Additional file 5: Sequence logos of predicted binding sites. The file contains sequence logos and their reverse complements of predicted binding sites inferred using the PFM(0), the PFM(1), and the PFM(2) for each of the 35 TFs. (PDF 11776 kb)

Additional file 6: ROC curves. The pdf file comprises for each TF one plot that shows the 25 ROC curves and one plot that shows the 25 PR curves from the 25–fold stratified repeated random sub-sampling validation procedure described in Methods 3. (PDF 2611.2 kb)

Abbreviations

MI: mutual information; PFM: phylogenetic footprinting model; PWM: position weight matrix; TF: transcription factor; TFBS: transcription factor binding site

Authors' contributions

MN and IG developed the key idea. MN and JC developed the computational methods. MN and HT performed the studies. All authors wrote, read, and approved the final manuscript.

Acknowledgements

We thank Ralf Eggeling, Jan Grau, Patrice Peterson, and Marcel Quint for valuable discussions. We thank the HudsonAlpha Institute for Biotechnology, the Stanford University, the Broad Institute of MIT and Harvard, and the University of Southern California for performing the ChIP-seq experiments and the ENCODE Analysis Working Group for providing the datasets.

Funding

This work was financially supported by DFG (grant no. GR3526/1), Gencat (2014 SGR 118), and Collectiveware (TIN2015-66863-C2-1-R).

Availability of data and materials

The datasets used in this work are included within the article and its additional files. The implementation of the proposed phylogenetic footprinting model is available at https://github.com/mgledi/PhyFoo/.

5. PREDICTING TRANSCRIPTION FACTOR BINDING SITES USING PHYLOGENETIC FOOTPRINTING

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle, Germany. [2]Leibniz Institute of Plant Biochemistry, Halle, Germany. [3]Institut d'Investigació en Intel·ligència Artificial, IIIA-CSIC, Campus UAB, Cerdanyola, Spain. [4]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany.

Received: 29 June 2016 Accepted: 24 January 2017
Published online: 01 March 2017

References

1. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013;14(3):204–20. doi:10.1038/nrg3354.
2. Tessarz P, Kouzarides T. Histone core modifications regulating nucleosome structure and dynamics. Nat Rev Mol Cell Biol. 2014;15(11):703–8. doi:10.1038/nrm3890.
3. Sainsbury S, Bernecky C, Cramer P. Structural basis of transcription initiation by RNA polymerase II. Nat Rev Mol Cell Biol. 2015;16(3):129–43. doi:10.1038/nrm3952.
4. Schoenberg DR, Maquat LE. Regulation of cytoplasmic mRNA decay. Nat Rev Genet. 2012;13(4):246–59. doi:10.1038/nrg3160.
5. de Fougerolles A, Vornlocher HP, Maraganore J, Lieberman J. Interfering with disease: a progress report on sirna-based therapeutics. Nat Rev Drug Discov. 2007;6(6):443–53.
6. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, et al. Pseudogene-derived small interfering rnas regulate gene expression in mouse oocytes. Nature. 2008;453(7194):534–8.
7. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008;321(5891):956–60.
8. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. Science. 2010;327(5968):996–1000.
9. Hobert O. Gene regulation by transcription factors and microrna. Science. 2008;319(5871):1785–6.
10. Voss TC, Hager GL. Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat Rev Genet. 2014;15(2):69–81.
11. Bulyk ML, Johnson PL, Church GM. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res. 2002;30(5):1255–61.
12. Salama RA, Stekel DJ. Inclusion of neighboring base interdependencies substantially improves genome-wide prokaryotic transcription factor binding site prediction. Nucleic Acids Res. 2010;38(12):135–5.
13. Eggeling R, Roos T, Myllymäki P, Grosse I. Inferring intra-motif dependencies of dna binding sites from chip-seq data. BMC Bioinforma. 2015;16(1):375.
14. Grau J, Posch S, Grosse I, Keilwagen J. A general approach for discriminative de novo motif discovery from high-throughput data. Nucleic Acids Res. 2013;41(21):e197. doi:10.1093/nar/gkt831.
15. Ma W, Noble WS, Bailey TL. Motif-based analysis of large nucleotide data sets using meme-chip. Nat Protoc. 2014;9(6):1428–50.
16. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8. doi:10.1038/nbt.3300.
17. Kulakovskiy I, Levitsky V, Oshchepkov D, Bryzgalov L, Vorontsov I, Makeev V. From binding motifs in chip-seq data to improved models of transcription factor binding sites. J Bioinforma Comput Biol. 2013;11(01):1340004.
18. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368–76.
19. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P, Odom DT. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Sci (New York, NY). 2010;328(5981):1036–40. doi:10.1126/science.1186176.
20. Hardison RC, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet. 2012;13(7):469–83.
21. Katara P, Grover A, Sharma V. Phylogenetic footprinting: a boost for microbial regulatory genomics. Protoplasma. 2012;249(4):901–7.
22. Martinez-Morales JR. Toward understanding the evolution of vertebrate gene regulatory networks: comparative genomics and epigenomic approaches. Brief Funct Genom. 2015. doi:10.1093/bfgp/elv032.
23. Blanchette M, Tompa M. Footprinter: a program designed for phylogenetic footprinting. Nucleic Acids Res. 2003;31(13):3840–2.
24. Sinha S, Blanchette M, Tompa M. Phyme: a probabilistic algorithm for finding motifs in sets of orthologous sequences. BMC Bioinforma. 2004;5(1):170.
25. Moses A, Chiang D, Pollard D, Iyer V, Eisen M. Monkey: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol. 2004;5(12):98. doi:10.1186/gb-2004-5-12-r98.
26. Neph S, Tompa M. Microfootprinter: a tool for phylogenetic footprinting in prokaryotic genomes. Nucleic Acids Res. 2006;34(suppl 2):366–8.
27. Newberg LA, Thompson WA, Conlan S, Smith TM, McCue LA, Lawrence CE. A phylogenetic gibbs sampler that yields centroid solutions for cis-regulatory site prediction. Bioinformatics. 2007;23(14):1718–27.
28. Siddharthan R, Siggia ED, Van Nimwegen E. Phylogibbs: a gibbs sampling motif finder that incorporates phylogeny. PLoS Comput Biol. 2005;1(7):67.
29. Siddharthan R. Phylogibbs-mp: module prediction and discriminative motif-finding by gibbs sampling. PLoS Comput Biol. 2008;4(8):1000156.
30. Arnold P, Erb I, Pachkov M, Molina N, van Nimwegen E. Motevo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of dna sequences. Bioinformatics. 2012;28(4):487–94. doi:10.1093/bioinformatics/btr695.
31. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74. doi:10.1038/nature11247.
32. UCSC. Genome Bioinformatics. 2016. http://hgdownload.cse.ucsc.edu/downloads.html. Accessed 29 Apr 2016.
33. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. Chip-seq guidelines and practices of the encode and modencode consortia. Genome Res. 2012;22(9):1813–31.
34. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotech. 2008;26(12):1351–9. doi:10.1038/nbt.1508.
35. ENCODE. Uniform TFBS composite track. http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/. Accessed 29 Apr 2016.
36. Multiple alignments of the hg19/GRCh37 human genome assembly. http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/. Accessed 29 Apr 2016.
37. 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):208–14.
38. Redhead E, Bailey TL. Discriminative motif discovery in dna and protein sequences using the deme algorithm. BMC Bioinforma. 2007;8(1):1.
39. Keilwagen J, Grau J, Paponov IA, Posch S, Strickert M, Grosse I. De-novo discovery of differentially abundant transcription factor binding sites including their positional preference. PLoS Comput Biol. 2011;7(2):1001070.
40. Agostini F, Cirillo D, Ponti RD, Tartaglia GG. Seamote: a method for high-throughput motif discovery in nucleic acid sequences. BMC Genomics. 2014;15(1):925.
41. Lawrence CE, Reilly AA. An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins Struct Funct Bioinforma. 1990;7(1):41–51.
42. Grau J, Keilwagen J, Gohr A, Haldemann B, Posch S, Grosse I. Jstacs: a java framework for statistical analysis and classification of biological sequences. J Mach Learn Res. 2012;13(1):1967–71.

5.3 Combining phylogenetic footprinting with motif models incorporating intra-motif dependencies

M Nettling, H Treutler, J Cerquides, I Grosse. 2017. Combining phylogenetic footprinting with motif models incorporating intra-motif dependencies. *BMC Bioinformatics*, 18:141 *doi: 0.1186/s12859-017-1495-1*

Bioinformatics, 2017, 1–8
doi: 10.1093/bioinformatics/btx033
Advance Access Publication Date: 27 January 2017
Original Paper

Phlogenetics

Unrealistic phylogenetic trees may improve phylogenetic footprinting

Martin Nettling[1,*], Hendrik Treutler[2], Jesus Cerquides[3] and Ivo Grosse[1,4]

[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle, Germany, [2]Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Halle, Germany, [3]Institut d'Investigació en Intel ligència Artificial, IIIA-CSIC, Campus UAB, Cerdanyola, Spain and [4]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on March 2, 2016; revised on December 2, 2016; editorial decision on January 18, 2017; accepted on January 19, 2017

Abstract

Motivation: The computational investigation of DNA binding motifs from binding sites is one of the classic tasks in bioinformatics and a prerequisite for understanding gene regulation as a whole. Due to the development of sequencing technologies and the increasing number of available genomes, approaches based on phylogenetic footprinting become increasingly attractive. Phylogenetic footprinting requires phylogenetic trees with attached substitution probabilities for quantifying the evolution of binding sites, but these trees and substitution probabilities are typically not known and cannot be estimated easily.

Results: Here, we investigate the influence of phylogenetic trees with different substitution probabilities on the classification performance of phylogenetic footprinting using synthetic and real data. For synthetic data we find that the classification performance is highest when the substitution probability used for phylogenetic footprinting is similar to that used for data generation. For real data, however, we typically find that the classification performance of phylogenetic footprinting surprisingly increases with increasing substitution probabilities and is often highest for unrealistically high substitution probabilities close to one. This finding suggests that choosing realistic model assumptions might not always yield optimal predictions in general and that choosing unrealistically high substitution probabilities close to one might actually improve the classification performance of phylogenetic footprinting.

Availability and Implementation: The proposed PF is implemented in JAVA and can be downloaded from https://github.com/mgledi/PhyFoo

Contact: martin.nettling@informatik.uni-halle.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene regulation is a highly complex process in nature based on several sub-processes such as transcriptional regulation including DNA methylation (Smith and Meissner, 2013), histone modifications (Tessarz and Kouzarides, 2014) and promotor escaping (Sainsbury *et al.*, 2015) as well as post-transcriptional regulation including modulated mRNA decay (Schoenberg and Maquat, 2012), siRNA interference (de Fougerolles *et al.*, 2007; Tam *et al.*, 2008) and alternative splicing (Luco *et al.*, 2010; Sultan *et al.*, 2008). One important step in this complex process is the regulation of transcriptional initiation by the interaction of transcription factors (TFs) with their binding sites (Hobert, 2008; Voss and Hager, 2014). Hence, identifying transcription factor binding sites (TFBSs) and inferring their binding motifs is a prerequisite in modern

biology, medicine and biodiversity research (Nowrousian, 2010; Villar *et al.*, 2014).

The last decade has witnessed a spectacular development of sequencing technologies unleashing new potentials in identifying TFBSs (Kulakovskiy *et al.*, 2010; Furey, 2012; Lasken and McLean, 2014; van Dijk *et al.*, 2016). Due to the increasing number of available genomes of different species and due to increasing computational resources, approaches for de-novo motif discovery based on phylogenetic footprinting have become increasingly attractive. Examples of highly popular tools for phylogenetic footprinting are *FootPrinter* (Blanchette and Tompa, 2003), *PhyME* (Sinha *et al.*, 2004), *MONKEY* (Moses *et al.*, 2004a), *PhyloGibbs* (Siddharthan *et al.*, 2005), *Phylogenetic Gibbs Sampler* (Newberg *et al.*, 2007), *PhyloGibbs-MP* (Siddharthan, 2008) and *MotEvo* (Arnold *et al.*, 2012). Supplementary Table S1 provides a comparison of these tools regarding the used evolutionary model, sequence model and learning principle.

One prerequisite for most phylogenetic footprinting approaches are multiple sequence alignments (MSAs) of upstream regions of orthologous genes of multiple not too closely related species (Anisimova *et al.*, 2013). These MSAs capture phylogenetic information, and the key idea of using these MSAs as starting point for phylogenetic footprinting results from the observations that (i) functional TFBSs are phylogenetically conserved and (ii) phylogenetically conserved TFBSs are aligned in MSAs. Examples of highly popular tools for aligning non-coding genomic regions are *T-Coffee* (Notredame *et al.*, 2000), *WebPRANK* (Löytynoja and Goldman, 2010) and *MAFFT* (Katoh and Standley, 2013).

Phylogenetic footprinting improves the de-novo motif discovery by incorporating phylogenetic dependencies within the MSA in contrast to approaches based on sequences from only one species. Substitution models of DNA sequence evolution such as the F81 model (Felsenstein, 1981) have been adapted to model the evolution of TFBSs in a position-specific manner, and it has been shown that these adapted models, which we call phylogenetic footprinting models (PFMs) for brevity, can detect TFBSs more accurately than models that neglect phylogenetic dependencies (Clark *et al.*, 2007; Gertz *et al.*, 2006; Hardison and Taylor, 2012; Hawkins *et al.*, 2009; Moses *et al.*, 2004a; Nettling *et al.*, 2017).

One fundamental prerequisite for phylogenetic footprinting is a phylogenetic tree including substitution probabilities attached to each of its branches, and choosing an appropriate phylogenetic tree and appropriate substitution probabilities is pivotal for the classification performance of phylogenetic footprinting (Kc and Livesay, 2011). However, estimating substitution probabilities within TFBSs is substantially harder than estimating them e.g. in protein-coding regions for at least two reasons:

First, the positions of TFBSs are unknown when performing phylogenetic footprinting, whereas the positions of protein-coding regions are known when estimating substitution probabilities there. Second, protein-coding regions are much longer than TFBSs, so one can use a much larger number of bases for estimating substitution probabilities for protein-coding regions than for TFBSs.

Estimating substitution probabilities within TFBSs is challenging, but several valuable studies have been performed in this direction (Doniger and Fay, 2007; Pollard *et al.*, 2010; Schaefke *et al.*, 2015; Tuğrul *et al.*, 2015). For example, studies on synthetic data have indicated that small substitution probabilities in the motif and moderate substitution probabilities in the flanking sequences can be preferable for motif recognition (Sinha *et al.*, 2004), and studies on different yeast species have confirmed these findings and shown that the likelihood of the Jukes-Cantor model (Jukes and Cantor, 1969) increases relative to a thymine background ('polyT') for small substitution probabilities in the motif and moderate substitution probabilities in the flanking sequences (Moses *et al.*, 2004b).

These and similar findings, however, have not lead to a robust approach of estimating substitution probabilities within TFBSs prior to or as part of phylogenetic footprinting, so the substitution probabilities are often simply taken from the literature or guessed, and their influence on the classification performance of phylogenetic footprinting has not yet been studied systematically.

Here, we study this influence based on a synthetic dataset and five real datasets of the TFs CTCF, GABP, NRSF, SRF and STAT1. Specifically, we describe the PFM, the datasets, the tested phylogenetic trees, the performance measure, and implementation details in section Methods, and we study the classification performance of phylogenetic footprinting as a function of the substitution rate for synthetic and real datasets, compare the results to those of phylogenetic footprinting based on expert trees from the literature, and discuss the findings in the context of several factors that affect the evolution of TFBSs in sections 3 and 4.

2 Materials and methods

In this section we describe (i) the used notation and the likelihood calculation of the PFM, (ii) the investigated datasets, (iii) the performance measure, (iv) the systematic investigation of phylogenetic trees and (v) the implementation of the PFMs.

2.1 Phylogenetic footprinting model

2.1.1 Notation

Our data contains N alignments, with each alignment containing O sequences (one per observed species) of length L_n.

Our phylogenetic model incorporates the existence of H additional *hidden* species, that is, species for which we cannot observe their sequences. Both hidden and observed species conform a tree. Thus, for each species k but the root, $pa(k)$ denotes the ancestor of species k in the tree. The root species is noted r.

Our probabilistic model contains a random variable $S_n^{u,k}$ for each nucleotide $1 \leq u \leq L_n$ of each species $1 \leq k \leq O+H$ of each alignment $1 \leq n \leq N$. These random variables take values in the set of bases $\mathcal{A} = \{A, C, G, T\}$. We note $\mathrm{pa}(S_n^{u,k})$ the uth nucleotide in the nth alignment of species $\mathrm{pa}(k)$ (the ancestor of k). By definition, the root has no ancestor and hence $\mathrm{pa}(S_n^{u,r}) = \varnothing$. We also refer to nucleotide $S_n^{u,k}$ as $A_n^{u,k}$ when species k is observed, and as $Y_n^{u,k}$ when species k is hidden. Furthermore we note by $Y_n^{u,\cdot}$ (respectively $S_n^{u,\cdot}$) the set containing each random variable $Y_n^{u,k}$ (respectively $S_n^{u,k}$), with $O+1 \leq k \leq O+H$ and Y_n the set containing every random variable in $Y_n^{u,\cdot}$ with $1 \leq u \leq L_n$.

An alignment A_n may or may not contain a TFBS. This is encoded in variable M_n, with M_n^0 indicating that alignment A_n does not contain a motif and M_n^1 indicating that alignment A_n does contain a motif.

2.1.2 Likelihood

The probability that the alignment A_n is generated by the PFM can be written as

$$p(A_n|\theta) = p(A_n|M_n^0, \theta) \times p(M_n^0|\theta) + p(A_n|M_n^1, \theta) \times p(M_n^1|\theta)$$

with variable M_n taking a Bernoulli distribution and θ denoting model parameters, namely the topology of the phylogenetic tree, the substitution probabilities and the evolutionary model with its

5. PREDICTING TRANSCRIPTION FACTOR BINDING SITES USING PHYLOGENETIC FOOTPRINTING

stationary probabilities for the flanking regions as well as the TFBS regions.

We need to specify the probability for non-motif-bearing $p(A_n|M_n^0,\theta)$ and for motif-bearing alignments $p(A_n|M_n^1,\theta)$. For reasons of clarity we omit θ in the following.

2.1.3 Likelihood of a non-motif-bearing alignment

The probability that alignment A_n is generated by the PFM as a non-motif bearing alignment is

$$p(A_n|M_n^0) = \sum_{Y_n} p(A_n|Y_n, M_n^0). \tag{1}$$

We assume that each single nucleotide alignment is independent of any other nucleotide alignment given θ and M_n^0. Furthermore, we assume that in each nucleotide alignment, the species satisfy the conditional independencies encoded by the phylogenetic tree. Thus,

$$p(A_n|M_n^0) = \prod_{u=1}^{L_n} \sum_{Y_n^{u\cdot}} p(S_n^{u\cdot}|M_n^0) \tag{2}$$

$$= \prod_{u=1}^{L_n} \sum_{Y_n^{u\cdot}} \prod_{k=1}^{O+H} p(S_n^{u,k}|\mathrm{pa}(S_n^{u,k}), M_n^0) \tag{3}$$

where

$$p(S_n^{u,k} = a|\mathrm{pa}(S_n^{u,k}) = b, M_n^0) = \begin{cases} \pi_0^a & \text{if } k = r \\ \gamma_k \times \pi_0^a + (1-\gamma_k)\delta_{a=b} & \text{if } k \neq r \end{cases}$$

according to the F81 model, where the base distribution of each position of the background sequence is denoted by π_0, the probability of a nucleotide a in the background sequence is denoted by π_0^a, and the substitution probability from the ancestor species to species k is denoted by γ_k. For more realistic phylogenetic models γ_k might also depend on specific nucleotide transitions.

2.1.4 Likelihood of a motif-bearing alignment

The probability that alignment A_n is generated by the PFM as a motif bearing alignment is

$$p(A_n|M_n^1) = \sum_{\ell_n=1}^{L_n-W+1} \sum_{Y_n} p(A_n, Y_n, \ell_n|M_n^1). \tag{4}$$

where W is the length of the TFBS and ℓ_n is the position of the TFBS in alignment A_n. Since single nucleotide alignments are assumed independent and considering the conditional independencies in the phylogenetic tree we have

$$p(A_n|M_n^1) = \sum_{\ell_n=1}^{L_n-W+1} p(\ell_n|M_n^1) \prod_{u=1}^{L_n} \sum_{Y_n^{u\cdot}} p(S_n^{u\cdot}|\ell_n, M_n^1) \tag{5}$$

with $p(S_n^{u\cdot}|\ell_n, M_n^1) = \prod_{k=1}^{O+H} p(S_n^{u,k}|\mathrm{pa}(S_n^{u,k}), \ell_n, M_n^1)$ and

$$p(S_n^{u,k}|pa(S_n^{u,k}), \ell_n, M_n^1) = \begin{cases} \pi_0^a & \text{if } k = r \text{ and } u < \ell_n \text{ or } u \geq \ell_n + W \\ \pi_{u-\ell_n+1}^a & \text{if } k = r \text{ and } \ell_n \leq u < \ell_n + W \\ \gamma_k \times \pi_0^a + (1-\gamma_k)\delta_{a=b} & \text{if } k \neq r \text{ and } u < \ell_n \text{ or } u \geq \ell_n + W \\ \gamma_k \times \pi_{u-\ell_n+1}^a + (1-\gamma_k)\delta_{a=b} & \text{if } k \neq r \text{ and } \ell_n \leq u < \ell_n + W \end{cases}$$

As for the non-motif-bearing alignment, the base distribution of each position of the background sequence is denoted by π_0 and the probability of a nucleotide a in the background sequence is denoted by π_0^a. The base distributions of a motif sequence of length W are denoted by π_w with $w \in [1,\ldots,W]$ and the probability of a nucleotide a at position w in a motif sequence is denoted by π_w^a. The substitution probability from the ancestor species to species k is denoted by γ_k.

Finally we assume motifs to be uniformly distributed, thus having that $p(\ell_n|M_n^1) = \frac{1}{L_n-W+1}$, which completes the specification of the likelihood function.

2.2 Data

2.2.1 Real data

The data used in this work originate from human ChIP-Seq data of the five TFs CTCF, GABP, NRSF, SRF and STAT1 Jothi *et al.* (2008); Valouev *et al.* (2008) and gapped alignments of the ChIP-Seq target regions from human with orthologous regions from monkey, cow, dog and horse. The original data provided by Arnold *et al.* (2012) consist of 900 gapped alignments for each of the five TFs. Each gapped alignment consists of sequences from six species. Since gapped alignments have a higher risk of showing mathematical side effects, we process them to derive ungapped alignments following three steps: (i) We remove the species that causes the highest number of gaps in all alignments. Accordingly, we remove sequences from opossum and keep orthologous regions from human, monkey, cow, dog and horse. (ii) In each alignment, we remove all alignment columns that contain at least one gap. (iii) We remove all alignments that are shorter than 21 bp, which is the length of the longest TFBS motif (NRSF) in the presented studies. Supplementary Table S2 shows details about the resulting datasets. All datasets are available as Supplementary Material.

2.2.2 Synthetic data

The synthetic dataset used in this work is generated using the PFM specified in section 2.1 with a star topology.

A negative set of 1000 non-motif-bearing alignments each of length $L = 300$ is generated. Each non-motif bearing alignment is generated in two steps as follows. (i) Sample the primordial sequence. For each position $u \in [1, L]$ of the sequence, sample a nucleotide from the uniform distribution π_0. (ii) For each of the descent species $o \in \{1,\ldots,5\}$, sample a mutated sequence given the primordial sequence position-wise. For each position $u \in [1, L]$, apply the F81 Felsenstein (1981) mutation model with the equilibrium distribution π_0 and substitution probability $\gamma = 0.2$ to the nucleotide of the primordial sequence at position u.

A positive set of 750 motif-bearing alignments each of length $L = 300$ is generated. Each motif-bearing alignment is generated as follows:

(i) Sample the primordial sequence given a TFBS length of $W = 15$.
   (a) Sample the start position $\ell \in [1, L - W + 1]$ of the TFBS from the uniform distribution.
   (b) For each position $u \in [1, \ell - 1]$ and $u \in [\ell + W, L]$ of the flanking sequence, we sample the nucleotide at position u from the uniform distribution π_0. For each position $u \in [\ell, \ell + W - 1]$ of the TFBS, we sample the nucleotide at position u from the distribution $\pi_{u-\ell_n+1}$. The distribution π_w with $w \in \{1,\ldots,15\}$ is uniformly drawn from the simplex.

(ii) For each of the descent species $o \in \{1,\ldots,5\}$, sample a mutated sequence given the primordial sequence position-wise.
   (a) For each position $u \in [1, \ell - 1]$ and $u \in [\ell + W, L]$ of the flanking sequence, apply the F81 mutation model with the equilibrium distribution π_0 and substitution probability $\gamma = 0.2$ to the nucleotide of the primordial sequence at position u.
   (b) For each position $u \in [\ell, \ell + W - 1]$ of the TFBS, apply the F81 mutation model with the equilibrium distribution $\pi_{u-\ell_n+1}$ and substitution probability $\gamma = 0.2$ to the nucleotide of the primordial sequence at position u.

2.3 Phylogenetic trees

To systematically investigate the influence of different phylogenetic trees on classification performance and hence on motif prediction, we introduce two simplifications. First, the underlying phylogenetic tree is a star topology implying that all species have one common ancestor. Second, all branches in the star topology have the same length, i.e. the probability that a base in the primordial sequence is replaced by a new base in a descendant sequence is the same for all sequences.

Now, it is possible to systematically vary the substitution probabilities $\gamma = \{0.05, 0.1, \ldots, 1.0\}$, where γ is inversely proportional to the phylogenetic relatedness. Small γ encode close phylogenetic relations and large γ encode distant phylogenetic relations. Especially, $\gamma = 1.0$ implies that the species are phylogenetically unrelated, i.e. the sequences of each alignment are statistically independent.

2.4 Classification performance

We evaluate all PFMs by a stratified repeated random sub-sampling validation by estimating all PFMs from a training set and measuring classification performance on a test set as follows.

In step 1, we generate two training sets and two disjoint test sets for each of the five TFs as follows. We randomly select 200 alignments from the set of alignments of a particular TF as positive training set, leaving the remaining alignments as positive test set. We perform a base shuffling on the positive set of alignments of the same TF to get a negative set of alignments. We randomly select 200 alignments from this set of alignments as negative training set and leave the remaining alignments as negative test set.

In step 2, we train a foreground model on the positive training set and a background model on the negative training set by expectation maximization (Lawrence and Reilly, 1990) using a numerical optimization procedure in the maximization step. We restart the expectation maximization algorithm, which is deterministic for a given dataset and a given initialization, 20 times with different initializations and choose the foreground model and the background model with the maximum likelihood on the positive training data and the negative training data, respectively, for classification. We use a likelihood-ratio classifier of the two chosen foreground and background models, apply this classifier to the disjoint positive and negative test sets, and calculate the area under the receiver operating characteristics curve and the area under the precision recall curve as measures of classification performance.

We repeat both steps 100 times and determine (i) the mean area under the receiver operating characteristic curve and its standard error and (ii) the mean area under the precision recall curve and its standard error.

2.5 Implementation

In order to investigate the influence of different phylogenetic trees in a fair and detailed way, we implement the proposed PFM based on the freely available Java Framework *Jstacs* (Grau *et al.*, 2012). Among others, *Jstacs* provides ready-to-use sequence models for reuse, numerical and non-numerical optimization procedures for model estimation, serialization of models and methods for the statistical evaluation of results. In contrast to existing tools which are typically focused on application, using *Jstacs* we are able to compare different PFMs in a detailed way by extracting mandatory information about the inferred models and the predicted TFBSs.

Algorithm 1 shows the pseudocode for inferring a PFM from a set of alignments. The implementation of the proposed PFM is available at https://github.com/mgledi/PhyFoo/.

Algorithm 1. Motif discovery algorithm for the proposed PFM. Upon random initialization of the model parameters we iteratively estimate sequence weights and model parameters in multiple algorithm restarts, where R denotes the number of restarts of the whole algorithm, and T denotes the number of iterations. The result is the set of model parameters together with maximum likelihood.

Data: Set of alignments $A = \{A_1, \ldots, A_N\}$
Flanking model: Maximize $p(A|\theta^1)$ for the model parameters $\pi_0 \subset \theta^1$
for r = 1 ...R **do**
  Initialize $\pi_w \subset \theta^1$ randomly for $w \in \{1, \ldots, W\}$
  for t = 1 ...T **do**
    E-step: Estimate $p(A_n|\ell_n, M_n^1, \theta^t)$ for each position ℓ_n in each alignment A_n given the model parameters θ^t
    M-step: Maximize the expected value of the complete-data log-likelihood with respect to model parameters π_w and denote the resulting argmax by θ^{t+1}.
  end for
  Keep θ^{T+1} denoted θ_r
end for
Result: $\theta \in \{\theta_1, \ldots \theta_R\}$ with maximum likelihood

3 Results

In this section, we investigate the classification performance of the PFM specified in section 2.1 as function of the substitution probability for a synthetic dataset and five real datasets. The synthetic dataset is generated using the PFM described in section 2.2. The five real datasets originate from human ChIP-Seq experiments of the five TFs CTCF, GABP, NRSF, SRF and STAT1 and MSAs of the predicted target regions with orthologous regions from monkey, cow, dog and horse as described in section 2.2.

In section 2.1.1, we study the likelihood of the popular PFM specified in section 2 as a function of the substitution probability for the synthetic dataset and the real dataset of TF CTCF. In section 2.1.2, we study the classification performance of the PFM as a function of the substitution probability for the same datasets. In section 2.1.3, we perform the studies of subsections 1 and 2 for the four datasets of the TFs GABP, NRSF, SRF and STAT1. In section 2.1.4, we study the classification performance of the PFM based on three selected phylogenetic trees for all five datasets of the TFs CTCF, GABP, NRSF, SRF and STAT1.

3.1 Likelihood on synthetic and real data

First, we test the implemented expectation maximization algorithm for the PFM specified in section 2.1 and summarized in **Algorithm 1** by applying it to synthetic data generated with a substitution probability of 0.2 as described in section 2.2 and to real data of TF CTCF. In both cases, we vary the substitution probability γ of the PFMs from 0.05 to 1.0 with increments of 0.05.

In case of synthetic data, we expect the best fit of the PFMs and thus the highest likelihood when the substitution probability γ of the PFMs is close to the substitution probability of 0.2 used for data generation. In case of real data of TF CTCF, we expect the best fit of the PFMs and thus the highest likelihood when the substitution

5. PREDICTING TRANSCRIPTION FACTOR BINDING SITES USING PHYLOGENETIC FOOTPRINTING

probability γ of the PFMs is in the range of $0.1 \leq \gamma \leq 0.4$ according to Gertz *et al.* (2006).

Figure 1a shows the likelihood as a function of the substitution probability γ ranging from 0.05 to 1.0 with increments of 0.05 for synthetic data, and we observe the expected function with a maximum at the substitution probability of $\gamma = 0.2$, which is equal to the substitution probability used for data generation. Figure 1b shows the likelihood as a function of the substitution probability γ for real data of TF CTCF, and we again observe the expected function with a maximum at the substitution probability of $\gamma = 0.2$, which is a reasonable value and in the range of $0.1 \leq \gamma \leq 0.4$ suggested by Gertz *et al.* (2006).

These findings indicate that the applied PFM and the applied maximum-likelihood principle are capable of identifying reasonable substitution probabilities for synthetic and real data of TF CTCF, where reasonable substitution probabilities mean substitution probabilities close to those used for data generation in case of synthetic data and in the range suggested by experts for real data of TF CTCF.

3.2 Classification performance on synthetic and real data

Second, we study the classification performance of the PFMs by the method described in section 2.3 on the same two datasets. We again vary γ from 0.05 to 1.0 with increments of 0.05 and compute the classification performance as a function of γ as described in section 2.4.

In case of both synthetic and real data, we expect that the classification performance looks qualitatively similar to the likelihood as a function of γ, i.e. we expect that the classification performance is highest for γ close to 0.2 for synthetic data and in the range of $0.1 \leq \gamma \leq 0.4$ for real data of TF CTCF.

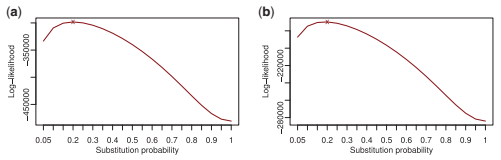


Fig. 1. Likelihood for different substitution probabilities. We plot the likelihood on synthetic data and CTCF data for a PFM using a star topology with all substitution probabilities set to $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$. **(a)** Synthetic data. Maximum likelihood is achieved for $\gamma = 0.2$, the substitution probability used for data generation. **(b)** CTCF data. Maximum likelihood is achieved for $\gamma = 0.2$, lying in the range of $0.1 \leq \gamma \leq 0.4$ suggested by the literature

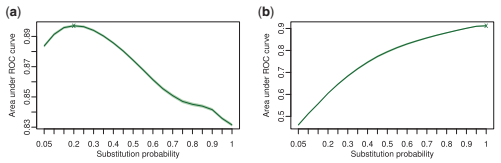


Fig. 2. Classification performance for different substitution probabilities. We plot the classification performance on synthetic data and CTCF data for a PFM using a star topology with all substitution probabilities set to $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$. **(a)** Synthetic data. Highest classification performance is achieved for $\gamma = 0.25$, which is close to $\gamma = 0.2$, the substitution probability used for data generation. **(b)** CTCF data. Highest classification performance is achieved for $\gamma = 1.0$, which is unrealistic and different from the expected result. We obtain similar results when quantifying the classification performance by the area under the PR curve (Supplementary Fig. S4)

Figure 2a shows the classification performance as a function of γ for synthetic data, and we observe the expected function with a maximum at $\gamma = 0.2$, which is equal to the substitution probability used for data generation and equal to the location of the maximum of the likelihood. These results are in agreement with those of Sinha *et al.* (2004) who additionally find that an underestimation of the true substitution probability leads to a more severe degradation of the classification performance than an overestimation of equal magnitude.

Figure 2b shows the classification performance as a function of γ for real data of TF CTCF, but here we observe a function that is different from the expected function, different from the function observed for synthetic data, and different from the likelihood function of Figure 1b. Specifically, we observe that the maximum is achieved for an unrealistically high value of $\gamma = 1.0$, which is clearly outside of the range of substitution probabilities of $0.1 \leq \gamma \leq 0.4$ suggested by Gertz *et al.* (2006) and much greater than the value of $\gamma = 0.2$ at which the maximum of the likelihood is located.

This observation is surprising because a substitution probability of $\gamma = 1.0$ corresponds to a PFM that assumes the orthologous sequences in the MSAs be statistically independent, i.e. phylogenetically unrelated. It indicates that choosing a realistic substitution probability in the range of $0.1 \leq \gamma \leq 0.4$ might lead to an inferior classification performance of phylogenetic footprinting compared to choosing an unrealistic substitution probability of $\gamma = 1.0$.

3.3 Classification performance and likelihood on four additional real datasets

Third, we study if the phenomenon that the maximum classification performance is achieved for an unrealistically high value of γ is specific for TF CTCF or possibly also present in other TFs. Hence, we perform the studies of sections 2.2.1 and 2.2.2 for four additional ChIP-Seq datasets of TFs GABP, NRSF, SRF and STAT1.

Figure 3a–d shows the four classification performances and the four likelihoods as functions of γ. For the likelihoods, we observe clear maxima for realistic substitution probabilities in the range of $0.1 \leq \gamma \leq 0.2$ in all four cases. However, for the classification performances, we observe the four maxima for unrealistically high substitution probabilities $\gamma \geq 0.8$. This observation is again surprising and states that the classification performance of phylogenetic footprinting is higher for an unrealistically high substitution probability of $\gamma = 1.0$ than for realistic substitution probabilities in the range of $0.1 \leq \gamma \leq 0.4$ for all five TFs CTCF, GABP, NRSF, SRF and STAT1.

In order to test if this result could be an artifact of the choice of the negative dataset, we study the classification performance when negatives are taken from the positives of the other datasets as done by Arnold *et al.* (2012). We obtain the same surprising results that the classification performance is higher for a substitution probability of $\gamma = 1.0$ than for realistic substitution probabilities for all five TFs (Supplementary Figs S5, S9, S13, S17 and S21).

Next, we scrutinize the motifs obtained by PFMs with a substitution probability of $\gamma = 1.0$. For synthetic data, we find that the motifs obtained by PFMs with $\gamma = 1.0$ are highly similar to the motifs used for data generation (Supplementary Fig. S1). For real data, we find that the motifs obtained by PFMs with $\gamma = 1.0$ are highly similar to the motifs obtained by PFMs with realistic substitution probabilities in the range of $0.1 \leq \gamma \leq 0.4$ (Supplementary Figs S2, S6, S10, S14 and S22). These findings suggest that the

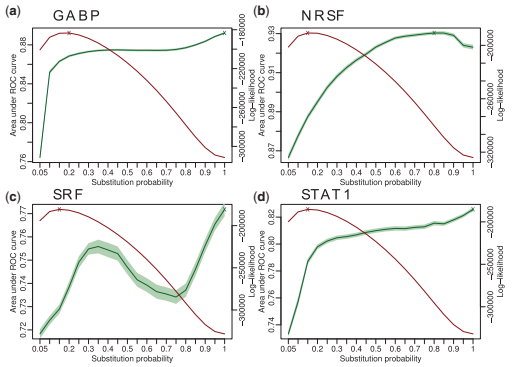


Fig. 3. Classification performance and likelihood for different substitution probabilities. We plot the classification performance (decreasing) and likelihood (increasing) on data of the four TFs GABP, NRSFm, SRF and STAT1 for substitution probabilities $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$. **(a)** GABP. The maximum likelihood is achieved for $\gamma = 0.2$. The best classification performance is achieved for $\gamma = 1.0$. **(b)** NRSF. Maximum likelihood is achieved for $\gamma = 0.15$. The best classification performance is achieved for $\gamma = 0.8$. **(c)** STAT1. The maximum likelihood is achieved for $\gamma = 0.15$. The best classification performance is achieved for $\gamma = 1.0$. **(d)** SRF. The maximum likelihood is achieved for $\gamma = 0.15$. The best classification performance is achieved for $\gamma = 1.0$. For each of the four TFs, we find qualitatively similar curves when quantifying the classification performance by the area under the PR curve (see Supplementary Figs S8, S12, S16 and S20)

motifs obtained by PFMs with an unrealistically high substitution probability of $\gamma = 1.0$ might be less biased than naively expected.

3.4 Classification performance using realistic phylogenetic trees

Fourth, we study if the phenomenon that the maximum classification performance is achieved for unrealistically high values of γ, which we observed for PFMs with a star topology, also occurs when using realistic phylogenetic trees. This study is motivated by observations that PFMs with phylogenetic trees with realistic tree topologies have the potential to yield higher classification performances than PFMs with phylogenetic trees with unrealistic star topologies (Newberg *et al.*, 2007; Palumbo and Newberg, 2010).

Hence, we study the classification performances of PFMs on synthetic data with different tree topologies and different substitution probabilities, and we find in all cases the highest classification performances near the substitution probabilities used for data generation (Supplementary Material section 4.2 and Supplementary Fig. S25). In addition to generating synthetic data by the F81 substitution model (Felsenstein, 1981), we also generate them by the more realistic HKY substitution model Hasegawa *et al.* (1985) in combination with different tree topologies and different substitution probabilities, and we find again the highest classification performances near the substitution probabilities used for data generation (Supplementary Material sections 4.4 and 4.5 and Supplementary Figs S27 and S28).

Next, we study the classification performance of the PFM on real data using a phylogenetic tree and substitution probabilities from the literature (Arnold *et al.*, 2012). We denote the PFM with a phylogenetic tree and substitution probabilities from the literature by $\mathcal{M}_{\text{lit}}^{\text{tree}}$, the PFM with a phylogenetic tree with a star topology and substitution probabilities according to the maximum-likelihood estimates of Figures 1b and 3a–d by $\mathcal{M}_{\text{ML}}^{\text{star}}$, and the PFM with a

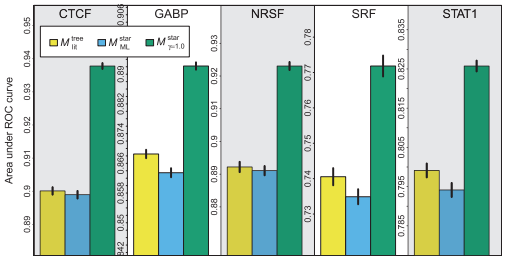


Fig. 4. Classification performance of three PFMs on real data of five TFs. The PFM $\mathcal{M}_{\gamma=1.0}^{\text{star}}$ (right) outperforms the PFMs $\mathcal{M}_{\text{lit}}^{\text{tree}}$ (left) and $\mathcal{M}_{\text{ML}}^{\text{star}}$ (middle), which implies that assuming phylogenetic independence generally improves motif prediction. The PFM $\mathcal{M}_{\text{lit}}^{\text{tree}}$ typically achieves a higher classification performance than the PFM $\mathcal{M}_{\text{ML}}^{\text{star}}$ (see Supplementary Table S3 for significances). For each of the five TFs, we find qualitatively similar results by the area under PR curve (see Supplementary Fig. S23) with similar significances shown in Supplementary Table S4. Supplementary Figures S23 also shows a comparison of $\mathcal{M}_{\gamma=1.0}^{\text{star}}$, $\mathcal{M}_{\text{ML}}^{\text{star}}$ and $\mathcal{M}_{\text{lit}}^{\text{tree}}$ with two additional PFMs

phylogenetic tree with a star topology and substitution probabilities of $\gamma = 1.0$ by $\mathcal{M}_{\gamma=1.0}^{\text{star}}$.

Figure 4 shows the classification performances of $\mathcal{M}_{\text{lit}}^{\text{tree}}$, $\mathcal{M}_{\text{ML}}^{\text{star}}$ and $\mathcal{M}_{\gamma=1.0}^{\text{star}}$ for each of the five TFs CTCF, GABP, NRSF, SRF and STAT1. Interestingly, we find that $\mathcal{M}_{\gamma=1.0}^{\text{star}}$ yields a significantly higher classification performance than the other two PFMs. In addition, we investigate the classification performances of PFMs with a star topology and a tree topology from the literature with branch lengths estimated from the data, and we find also in this case that $\mathcal{M}_{\gamma=1.0}^{\text{star}}$ yields a significantly higher classification performance than the other two PFMs (Supplementary Material section 3 and Supplementary Fig. S23).

These findings state that, in case of real data, choosing unrealistic model assumptions—namely a phylogenetic tree with a star topology and substitution probabilities of $\gamma = 1.0$—might yield higher classification performances than the same PFMs with more realistic phylogenetic trees and more realistic substitution probabilities.

4 Discussion

Possible explanations for this unexpected observation might be unrealistic model assumptions of the substitution model, heterogeneous substitution probabilities at different TFBS positions and in different DNA regions, heterotachious substitution probabilities at different times of evolution, or the construction of incorrect or at least partially erroneous MSAs.

Violations of model assumptions sometimes lead to a poor classification performance or to a strange dependence of the classification performance on one or several model parameters. Such a situation might occur in phylogenetic footprinting, where PFMs typically assume the same phylogenetic tree and the same substitution probabilities for all positions of all TFBSs, for all TFBSs and all of their flanking regions, and for all chromosomal regions and all MSAs despite the fact that all of these assumptions are almost certainly violated (Conrad *et al.*, 2011; Lercher and Hurst, 2002; Moses *et al.*, 2003; Schuster-Böckler and Lehner, 2012; Tian *et al.*, 2008; Weber *et al.*, 2007; Wolfe *et al.*, 1989).

Heterogeneous substitution probabilities among different DNA regions are omnipresent and typically taken into account when modeling the evolution of proteins or protein-coding genes. However, this heterogeneity is typically neglected in PFMs, where this

assumption would lead to potential over-fitting (Hawkins, 2004) due to the facts that the positions of TFBSs are unknown in phylogenetic footprinting and that TFBSs are much shorter than protein-coding genes.

Heterotachious substitution probabilities, i.e., substitution probabilities that vary with time, are another feature that is typically neglected in PFMs despite being omnipresent in both functional TFBSs as well as their flanking regions. Neglecting heterotachy might lead to the estimation of severely biased substitution probabilities, to incorrect motif predictions, and thus to a poor classification performance (Kolaczkowski and Thornton, 2004).

Incorrect or at least partially erroneous MSAs are another problem that might lead to the violation of model assumptions (Kim and Ma, 2011; Löytynoja *et al.*, 2012). In particular, insertions and deletions as well as heterogeneity in sequence composition such as a varying GC-content (Hardison and Taylor, 2012) might cause MSA algorithms to become potentially imprecise and might thus affect all downstream analyses (Löytynoja and Goldman, 2008).

Maximum-likelihood estimators can be proven to achieve the highest classification performance in the asymptotic limit of infinitely large datasets and under the prerequisite that the models used for classification are exactly those used for data generation. However, both prerequisites are typically not fulfilled in practice, so it often happens that the highest classification performance is not achieved by those parameters that maximize the likelihood.

This situation apparently occurs for phylogenetic footprinting in a surprisingly pronounced manner, which seems to indicate that the likelihoods of currently used PFMs are less affected by violated model assumptions than their classification performances. On an intuitive level, PFMs with realistic phylogenetic trees and realistic substitution probabilities seem to be more strongly affected by heterogeneity, heterotachy and errors in MSAs than PFMs with unrealistically high substitution probabilities, so using such unrealistically high substitution probabilities might by a temporarily useful choice until more sophisticated PFMs capable of coping with heterogeneity, heterotachy and errors in MSAs are being developed.

5 Conclusions

We have studied the influence of choosing different phylogenetic trees and different substitution probabilities on the likelihood and the classification performance of PFMs. We have performed these studies on synthetic and real data obtained from ChIP-Seq experiments performed in human and MSAs of ChIP-Seq positive regions with upstream regions of orthologous genes in monkey, cow, dog and horse.

We find that the likelihood depends on the substitution probability in a qualitatively similar manner for synthetic and real data, where it reaches a maximum for realistic substitution probabilities in the range of $0.1 \leq \gamma \leq 0.2$. In contrast, we find that the classification performance depends on the substitution probability in a qualitatively different manner for synthetic and real data.

For synthetic data, the classification performance reaches a maximum at the values of the substitution probability used for data generation, which coincide with those values that maximize the likelihood. For real data, however, it increases with the substitution probability and stops increasing only at unrealistically high values of the substitution probability in the range of $0.8 \leq \gamma \leq 1$, which are very different from those values that maximize the likelihood.

We find in all of the studied datasets that PFMs using unrealistic substitution probabilities of $\gamma = 1.0$ yield higher classification performances than PFMs using realistic substitution probabilities.

One possible explanation for this strange behavior of the classification performance on the substitution probability is the presence of heterogeneous and heterotachious substitution probabilities, which are neglected by currently used PFMs, and the sensitive dependence of PFMs on the reconstructed MSAs that might be partially incorrect.

Apparently, PFMs using unrealistic substitution probabilities of $\gamma = 1.0$ are more robust to these and possibly other violations of the model assumptions than PFMs based on realistic substitution probabilities, and this robustness might lead to less biased parameter estimates and thus more accurate phylogenetic footprints.

This observation leads to the strange practical recommendation of using PFMs using unrealistic substitution probabilities of $\gamma = 1.0$ instead of using PFMs using realistic substitution probabilities until there are more sophisticated models for the evolution of TFBSs and their flanking regions that take into account heterogeneity and heterotachy as well as partially erroneous alignments in a position-specific manner.

Acknowledgements

We thank Karin Breunig, Ralf Eggeling, Jan Grau, and Peter Stadler for valuable discussions.

Funding

We thank DFG [grant no. GR3526/1] for financial support.

Conflict of Interest: none declared.

References

Anisimova,M. *et al.* (2013) State-of the art methodologies dictate new standards for phylogenetic analysis. *BMC Evolution. Biol.*, **13**, 161.

Arnold,P. *et al.* (2012) Motevo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of dna sequences. *Bioinformatics*, **28**, 487–494.

Blanchette,M. and Tompa,M. (2003) Footprinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res.*, **31**, 3840–3842.

Clark,A.G. *et al.* (2007) Evolution of genes and genomes on the drosophila phylogeny. *Nature*, **450**, 203–218.

Conrad,D.F. *et al.* (2011) Variation in genome-wide mutation rates within and between human families. *Nature*, **43**, 712–714.

de Fougerolles,A. *et al.* (2007) Interfering with disease: a progress report on sirna-based therapeutics. *Nat. Rev. Drug Discov.*, **6**, 443–453.

Doniger,S.W. and Fay,J.C. (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput. Biol.*, **3**, e99.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Furey,T.S. (2012) ChIPseq and beyond: new and improved methodologies to detect and characterize proteinDNA interactions. *Nat. Rev. Genet.*, **13**, 840–852.

Gertz,J. *et al.* (2006) Phylogeny based discovery of regulatory elements. *BMC Bioinformatics*, **7**, 266.

Grau,J. *et al.* (2012) Jstacs: a java framework for statistical analysis and classification of biological sequences. *J. Mach. Learn. Res.*, **13**, 1967–1971.

Hardison,R.C. and Taylor,J. (2012) Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.*, **13**, 469–483.

Hasegawa,M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J. Mol. Evol.*, **22**, 160–174.

Hawkins,D.M. (2004) The problem of overfitting. *J. Chem. Inform. Comput. Sci.*, **44**, 1–12.

Hawkins,J. *et al.* (2009) Assessing phylogenetic motif models for predicting transcription factor binding sites. *Bioinformatics*, **25**, i339–i347.

Hobert,O. (2008) Gene regulation by transcription factors and micrornas. *Science*, **319**, 1785–1786.

Jothi,R. *et al.* (2008) Genome-wide identification of in vivo protein-dna binding sites from chip-seq data. *Nucl. Acids Res.*, **36**, 5221–5231.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. *Mammal. Protein Metab.*, **3**, 132.

Katoh,K. and Standley,D.M. (2013) Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Kc,D.B. and Livesay,D.R. (2011) Topology improves phylogenetic motif functional site predictions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. **8**, 226–233.

Kim,J. and Ma,J. (2011) Psar: measuring multiple sequence alignment reliability by probabilistic sampling. *Nucleic Acids Res.*, **39**, 6359–6368.

Kolaczkowski,B. and Thornton,J.W. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, **431**, 980–984.

Kulakovskiy,I.V. *et al.* (2010) Deep and wide digging for binding motifs in chip-seq data. *Bioinformatics*, **26**, 2622–2623.

Lasken,R.S. and McLean,J.S. (2014) Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat. Rev. Genet.*, **15**, 577–584.

Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.

Lercher,M.J. and Hurst,L.D. (2002) Human snp variability and mutation rate are higher in regions of high recombination. *Trends Genet.*, **18**, 337–340.

Löytynoja,A. and Goldman,N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632–1635.

Löytynoja,A. and Goldman,N. (2010) webprank: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*, **11**, 579.

Löytynoja,A. *et al.* (2012) Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics*, **28**, 1684–1691.

Luco,R.F. *et al.* (2010) Regulation of alternative splicing by histone modifications. *Science*, **327**, 996–1000.

Moses,A.M. *et al.* (2004a) Monkey: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.*, **5**, R98.

Moses,A.M. *et al.* (2003) Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol. Biol.*, **3**, 19.

Moses,A.M. *et al.* (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pacific Symposium on Biocomputing*. Hawaii, United States, pp. 324–335.

Nettling,M. *et al.* (2017) Combining phylogenetic footprinting with motif models incorporating intra-motif dependencies. *BMC Bioinformatics* (In press).

Newberg,L.A. *et al.* (2007) A phylogenetic gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics*, **23**, 1718–1727.

Notredame,C. *et al.* (2000) T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Nowrousian,M. (2010) Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryot. Cell*, **9**, 1300–1310.

Palumbo,M.J. and Newberg,L.A. (2010) Phyloscan: locating transcription-regulating binding sites in mixed aligned and unaligned sequence data. *Nucleic Acids Res.*, **38**, W268–W274.

Pollard,K.S. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.

Sainsbury,S. *et al.* (2015) Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.*, **16**, 129–143.

Schaefke,B. *et al.* (2015) Gains and losses of transcription factor binding sites in saccharomyces cerevisiae and saccharomyces paradoxus. *Genome Biol. Evol.*, **7**, 2245–2257.

Schoenberg,D.R. and Maquat,L.E. (2012) Regulation of cytoplasmic mRNA decay. *Nat. Rev.. Genet.*, **13**, 246–259.

Schuster-Böckler,B. and Lehner,B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, **488**, 504–507.

Siddharthan,R. (2008) Phylogibbs-mp: module prediction and discriminative motif-finding by gibbs sampling. *PLoS Comput. Biol.*, **4**, e1000156.

Siddharthan,R. *et al.* (2005) PhyloGibbs: a gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67.

Sinha,S. *et al.* (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, **5**, 170.

Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat. Rev. Genet.*, **14**, 204–220.

Sultan,M. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.

Tam,O.H. *et al.* (2008) Pseudogene-derived small interfering rnas regulate gene expression in mouse oocytes. *Nature*, **453**, 534–538.

Tessarz,P. and Kouzarides,T. (2014) Histone core modifications regulating nucleosome structure and dynamics. *Nat. Rev. Mol. Cell Biol.*, **15**, 703–708.

Tian,D. *et al.* (2008) Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature*, **455**, 105–108.

Tuğrul,M. *et al.* (2015) Dynamics of transcription factor binding site evolution. *PLoS Genet.*, **11**, e1005639.

Valouev,A. *et al.* (2008) Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat. Methods*, **5**, 829–834.

van Dijk,E.L. *et al.* (2016) Ten years of next-generation sequencing technology. *Trends Genet.*, **30**, 418–426.

Villar,D. *et al.* (2014) Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat. Rev. Genet.*, **15**, 221–233.

Voss,T.C. and Hager,G.L. (2014) Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat. Rev. Genet.*, **15**, 69–81.

Weber,M. *et al.* (2007) Distribution, silencing potential and evolutionary impact of promoter dna methylation in the human genome. *Nat. Genet.*, **39**, 457–466.

Wolfe,K.H. *et al.* (1989) Mutation rates differ among regions of the mammalian genome. *Nature*, 283–285. pages

6 Visualisation of motifs

6.1 DiffLogo: A comparative visualisation of sequence motifs

M Nettling*, H Treutler*, J Grau, J Keilwagen, S Posch, I Grosse. 2015. DiffLogo: a comparative visualization of sequence motifs. *BMC bioinformatics*, 16:1 *doi:10.1186/s12859-015-0767-x*

Nettling *et al. BMC Bioinformatics* (2015) 16:387
DOI 10.1186/s12859-015-0767-x

BMC Bioinformatics

SOFTWARE Open Access

DiffLogo: a comparative visualization of sequence motifs

Martin Nettling[1*†], Hendrik Treutler[2†], Jan Grau[1], Jens Keilwagen[3], Stefan Posch[1] and Ivo Grosse[1,4]

Abstract

Background: For three decades, sequence logos are the *de facto* standard for the visualization of sequence motifs in biology and bioinformatics. Reasons for this success story are their simplicity and clarity. The number of inferred and published motifs grows with the number of data sets and motif extraction algorithms. Hence, it becomes more and more important to perceive differences between motifs. However, motif differences are hard to detect from individual sequence logos in case of multiple motifs for one transcription factor, highly similar binding motifs of different transcription factors, or multiple motifs for one protein domain.

Results: Here, we present *DiffLogo*, a freely available, extensible, and user-friendly R package for visualizing motif differences. *DiffLogo* is capable of showing differences between DNA motifs as well as protein motifs in a pair-wise manner resulting in publication-ready figures. In case of more than two motifs, *DiffLogo* is capable of visualizing pair-wise differences in a tabular form. Here, the motifs are ordered by similarity, and the difference logos are colored for clarity. We demonstrate the benefit of *DiffLogo* on CTCF motifs from different human cell lines, on E-box motifs of three basic helix-loop-helix transcription factors as examples for comparison of DNA motifs, and on F-box domains from three different families as example for comparison of protein motifs.

Conclusions: *DiffLogo* provides an intuitive visualization of motif differences. It enables the illustration and investigation of differences between highly similar motifs such as binding patterns of transcription factors for different cell types, treatments, and algorithmic approaches.

Keywords: Sequence analysis, Sequence logo, Sequence motif, Position weight matrix, Binding sites

Background

Biological polymer sequences encode information by the order of their monomers, i.e., bases or amino acids. Often specific parts of the polymer sequence are of particular interest, as they encode, for instance, the binding of transcription factors to specific binding sites [1, 2], the binding to micro-RNA-targets in mRNAs, splice donor sites and splice acceptor sites in pre-mRNAs [3, 4], the presence of phosphorylation sites in proteins, or the folding of specific protein domains [5]. The set of subsequences of one specific biological process are often represented as a sequence motif.

A sequence motif is a model, that represents the preference for the monomers based on a set of aligned biopolymer sequences. Sequence motifs are the result of pipelines comprising wet-lab experiments and motif prediction algorithms, and are frequently used as the basis of *in silico* predictions [6]. Thus, sequence motif are critical for research of a wide range of problems in biology and bioinformatics.

Considering a particular transcription factor, there are many pipelines that combine wet-lab experiments such as *HT-SELEX* [7, 8], *ChIP-Seq* [9] or *DNase-Seq footprinting* [10] with motif prediction algorithms such as *MEME* [2, 11], *ChIPMunk* [12], *POSMO* [13], or *Dimont* [14]. Wet-lab experiments differ in their experimental setup, e.g., ecotypes, cell types, developmental stage, time points, or treatment, and motif prediction algorithms differ in their mathematical theory and implementation details.

Visualizing the results of motif discovery is nowadays accomplished by sequence logos [15], the *de facto*

*Correspondence: martin.nettling@informatik.uni-halle.de
†Equal contributors
[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
Full list of author information is available at the end of the article

standard for visualizing motifs in biology and bioinformatics. Sequence logos emerged as an essential tool for researchers to interpret findings, document work, share knowledge, and present results.

However, comparing multiple sequence logos by visual inspection is sometimes tricky. Differences between sequence logos of two unrelated transcription factors are usually obvious, whereas differences between sequence logos of the same transcription factor are often less obvious and rather hard to perceive as depicted in Fig. 1. Moreover, the results of motif discovery algorithms need to be compared against huge reference databases such as *JASPAR* [16] or *UniProbe* [17] or motifs from literature.

For this reason, the comparison of motifs is of primary interest. Several numerical measures including variants of Euclidean distance, Pearson correlation, and Jensen-Shannon divergence have been used to compare motifs [18–21]. These measures express the difference of motifs as a single number that can be easily utilized subsequently, e.g., for rankings or clustering algorithms. However, these measures lose the information of what exactly makes the difference between the motifs of interest. Hence, the comparison of multiple pairs of motifs can result in similar measures.

There are various tools for the analysis and visualization of motifs as summarized in Table 1. The R package *seqLogo* [22] is an implementation of sequence logos. In the context of motif comparison, sequence logos may be interpreted as a comparison of the input motif with a uniformly distributed motif. The web application *iceLogo* [23] extends this approach by comparing the input motif with a motif that follows the same background distribution at each motif position. Basically, *seqLogo* and *iceLogo* are designed for the presentation of single motifs. In contrast, the R package *MotifStack* [24] and the web application *STAMP* [25] are designed for the presentation of multiple motifs. Here, the input motifs are clustered and presented as sequence logos. Thus, the approach of both tools may be interpreted as multiple comparisons with a uniformly distributed motif. The web application *Two Sample Logo* [26] is capable of comparing two input motifs on the basis of probability theory. This comparison is performed for each motif position individually and results in a sophisticated motif comparison. Depending on the focus of each tool, the input format is a set of aligned sequences and/or a position frequency matrix or position weight matrix. In addition, some tools focus exclusively on DNA motifs, while others cover DNA, RNA, and protein motifs or even allow arbitrary alphabets. Table 1 summarizes tools and their capabilities. In section 4 of Additional file 1, we additionally provide comparative example plots generated by *seqLogo*, *iceLogo*, *STAMP*, *Two Sample Logo*, and *DiffLogo*.

We intend the pair-wise comparison of motifs and extend this idea towards the comparison of multiple motifs as follows.

We focus on the comparison of position-specific symbol distributions of two motifs. We neglect dependencies between different motif positions to reduce complexity. As suggested by the *sequence logo* approach, we intend to represent the characteristics of each motif position by the two properties stack height and symbol height within a stack. The stack height is to be proportional to the degree of distribution dissimilarity. The symbol height is to be proportional to the degree of differential symbol abundance.

We intend to compare three or more motifs on the basis of pair-wise motif comparisons. This comparison is to take into account all pair-wise motif comparisons, suggesting an arrangement in a grid with one row and one column for each motif and one cell for each motif comparison. Similar motifs are to be placed in nearby rows and columns, and the degree of similarity between all motifs is to become obvious at a glance analogous to heatmaps. The grid is to be complemented with a display of the individual sequence logos for further comparisons.

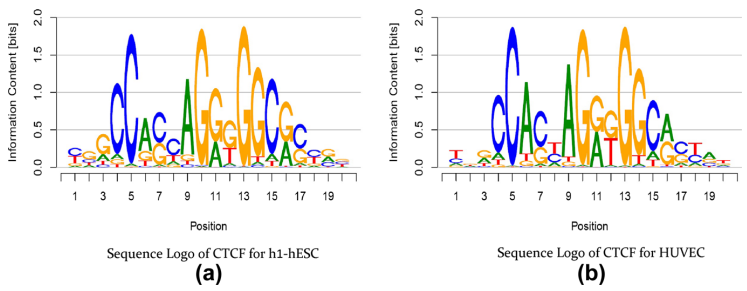


Fig. 1 Sequence logos of CTCF motifs from cell lines H1-hESC and HUVEC. The two sequence logos are highly similar in their conservation profile (height of stacks) and nucleotide preference at the individual motif positions

Table 1 Comparison of related tools. We compare six publicly available tools on the basis of five criteria

	Features				
Tools	Alphabet	Input format	Comparison	Clustering	Extensible
seqLogo	DNA	matrix	uniform	-	-
iceLogo	DNA/RNA, proteins	sequences	average	-	-
MotifStack	any	matrix	uniform	hclust	-
STAMP	DNA	sequences, matrix	uniform	UPGMA/SOTA	-
Two Sample Logo	DNA/RNA, proteins	sequences	position-specific	-	-
DiffLogo	any	sequences, matrix	position-specific	hclust, optimal leaf ordering	✓

In the first and second column, we examine the kind of supported input, in the third and forth column we examine the mode of action, and in the fifth column we examine whether the tool is extensible. For the criterion "alphabets" we summarize the supported biopolymers out of DNA, RNA, and proteins or arbitrary alphabets in case of "any". For the criterion "input format" we discriminate a set of "sequences" versus "matrix", which addresses at least one out of the formats position weight matrix (PWM), position frequency matrix (PFM), and position count matrix (PCM). For the criterion "comparison" we characterize the kind of distribution that is used for motif comparison ("uniform" is the uniform distribution, "average" is the average base distribution in a set of sequences, and "position-specific" is a position-specific distribution). For the criterion "clustering" we point out whether there is a clustering of motifs and which cluster-algorithm is used. For the criterion "extensible" we note whether the tool is extensible by the user

Implementation

In this section, we first define the used notation. We then briefly describe the classical sequence logo. Subsequently, we introduce the difference logo for the visualization of pair-wise motif differences. We discuss this new method and explore potential biological interpretations. Finally, we propose an approach for employing difference logos for the joint comparison of multiple motifs.

Basic notation and sequence logo

Consider a motif as an abstract description of a given set of aligned sequences of common length L from the alphabet $\mathcal{A}$. The relative frequency of symbol $a \in \mathcal{A}$ at position $\ell \in [1, L]$ corresponds to the (estimated) probability $p_{\ell,a}$. In case of two motifs, we use $p_{\ell,a}$ for the first motif and analogously $q_{\ell,a}$ for the second motif.

The well-known sequence logo visualizes a motif with a symbol stack for each position. We denote the height of the stack at position ℓ by H_ℓ and the height of symbol a within this stack by $H_{\ell,a}$. In the traditional sequence logo, H_ℓ and $H_{\ell,a}$ are defined by

$$H_\ell = \log_2(|\mathcal{A}|) - \sum_{a\in\mathcal{A}} p_{\ell,a} \cdot \log_2(p_{\ell,a}) \quad (1)$$

$$H_{\ell,a} = p_{\ell,a} \cdot H_\ell, \quad (2)$$

which states that the height of a stack at position ℓ reflects the degree of conservation at position ℓ quantified by the information content and that the height of each symbol at position ℓ is proportional to its frequency at position ℓ. Hence, the traditional sequence logo is an intuitive visualization of both (i) conserved motif positions and (ii) abundant bases.

The approach of *DiffLogo*

As specified earlier, we compare motifs per position. Similar to the sequence logo, we show a symbol stack for each position. We redefine the calculation of H_ℓ and use this measure as the total height of position ℓ reflecting the difference of the symbol distribution of both motifs at this position. We redefine the calculation of $H_{\ell,a}$ and use this measure as the height of a symbol within the stack at position ℓ. In the following, $H_{\ell,a}$ can be positive or negative. Symbols with positive values $H_{\ell,a}$ are plotted upward. Symbols with negative values $H_{\ell,a}$ are plotted downward.

Generally, there is a plethora of well-understood mathematical criteria that can be combined to define the height of a symbol stack and the relative heights of symbols within the stack such as probability differences, information divergences, distance measures, or entropies [27]. In the following, we present *DiffLogo* with the example of the Jensen-Shannon divergence for the calculation of H_ℓ and normalized probability differences for the calculation of $H_{\ell,a}$. We denote the combination of these two measures as weighted difference of probabilities.

Weighted difference of probabilities

We calculate the stack height for each motif position using the Jensen-Shannon divergence. The Jensen-Shannon divergence is a measure for the dissimilarity of two probability distributions based on information theory [28] (see Fig. 2). In contrast to other measures, the Jensen-Shannon divergence shows a comparable behavior when evaluating dissimilarities of distributions near the uniform distribution. The Jensen-Shannon divergence of two motifs at position ℓ is given by

$$H_\ell = \frac{1}{2}\sum_{a\in\mathcal{A}} p_{\ell,a} \log_2 \frac{p_{\ell,a}}{m_{\ell,a}} + \frac{1}{2}\sum_{a\in\mathcal{A}} q_{\ell,a} \log_2 \frac{q_{\ell,a}}{m_{\ell,a}}, \quad (3)$$

where $m_{\ell,a} = \frac{p_{\ell,a}+q_{\ell,a}}{2}$.

We define the height of each symbol by

$$H_{\ell,a} = r_{\ell,a} \cdot H_\ell, \quad (4)$$

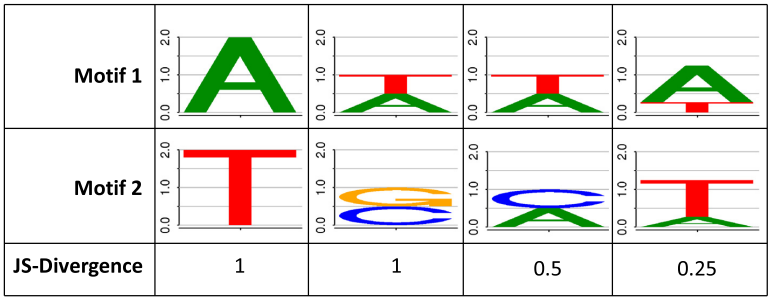


Fig. 2 Exemplary comparison of four DNA motifs of length one using the Jensen-Shannon divergence. Motif 1 and motif 2 are depicted as sequence logos. For each column, we compare the motif in the first row with the motif in the second row using the Jensen-Shannon divergence listed in the third row. In the first example we depict the case with only one base in each motif resulting in a maximal Jensen-Shannon divergence of 1 bit. In the second example we depict the case with two equally abundant bases both in motif 1 and motif 2 (both different) resulting again in a maximal Jensen-Shannon divergence of 1 bit. In the third example we depict the case with two equally abundant bases both in motif 1 and motif 2 (one equal and one different) resulting in a Jensen-Shannon divergence of 0.5 bit. In the fourth example we depict the case with two bases both in motif 1 and motif 2 (differentially abundant) resulting in a Jensen-Shannon divergence of 0.25 bit

where we define the weight $r_{\ell,a}$ as

$$r_{\ell,a} = \begin{cases} \frac{p_{\ell,a} - q_{\ell,a}}{\sum_{a' \in \mathcal{A}} |p_{\ell,a'} - q_{\ell,a'}|} & \text{if } p_\ell \neq q_\ell \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

$r_{\ell,a}$ is the probability difference of symbol a at position ℓ between two motifs normalized by the sum of absolute probability differences at this position. We use normalized probability differences as these are indicators for the gain or loss of symbol abundance and provide a view on the symbol distribution differences of both motifs. As a consequence, symbols less abundant in the second motif compared to the first motif are plotted upward, and symbols more abundant in the second motif compared to the first motif are plotted downward.

This representation emphasizes a high gain or loss of probability in co–occurrence with a high gain or loss of information content. The sum of the heights of symbols with a gain of probability and the sum of the heights of symbols with a loss of probability are equal at every position, because each gain of probability of one symbol implies a loss of probability of the remaining symbols. The advantage of this approach is that we are capable of seeing differences of position-specific symbol distributions and of seeing those symbols that are responsible for these differences by gaining or losing abundance.

Comparison of multiple motifs

According to the requirements formulated above, we propose a visualization for the joint comparison of $N \geq 3$ motifs given the measure H_ℓ as follows.

We plot the difference logos of all $N \times (N-1)$ motif pairs with a common ordinate scaling. We define a scalar dissimilarity value D for a pair of motifs as the sum of all stack heights in the corresponding difference logos,

$$D = \sum_{\ell=1}^{L} H_\ell. \quad (6)$$

We compute a motif order to group similar motifs. Here, we take the optimal leaf order of a hierarchical clustering of the motifs based on D (function *hclust* in R package *stats* and function *order.optimal* in R package *cba*). We arrange the difference logos ordered in an $N \times N$ grid with an empty diagonal. Difference logos opposing each other across the diagonal of the grid correspond to each other by an inversion of the ordinate. We visualize D with the background color of the corresponding difference logo using a color gradient from green (most similar among all pairwise comparisons) to red (most dissimilar). We outline the motif names above each column and left of each row. In addition, we allow the possibility of drawing the classic sequence logos and the cluster tree above the columns as auxiliary information.

The advantage of this approach is that we are capable of surveying the overall similarities and dissimilarities in the resulting difference logo grid. Greenish regions indicate similar motif groups and reddish rows and columns indicate less similar motifs. Given a region of interest, it is furthermore possible to comprehend the origins of dissimilarities from the individual difference logos and optionally the sequence logos.

R package

DiffLogo is written in R [29]. We provide the implementation as a ready-to-use R package. For symbol drawing, *DiffLogo* uses adapted methods from the package

seqLogo [22] in the software suite *bioconductor* [30]. *DiffLogo* allows the analysis of sequence motifs defined over arbitrary alphabets.

The core functions can be parameterized with functions for H_ℓ and $r_{\ell,a}$. Hence, the user is capable of combining different formulae for H_ℓ and $r_{\ell,a}$. We provide implementations of the Jensen-Shannon divergence and the normalized probability difference used for the difference logos presented in this manuscript. In addition, *DiffLogo* provides other implementations for H_ℓ and $r_{\ell,a}$ as alternatives. Exemplarily, we show the result of eight different combinations of measures for stack height and symbol height in Additional file 1: Tables S1 and S2. The *DiffLogo* package comprises example data, example code, and further documentation.

Results and discussion

In this section, we present three examples demonstrating the utility of *DiffLogo* in different applications. First, we examine differences in motifs of DNA binding sites of the same transcription factor from five different cell lines. Second, we examine differences in motifs of DNA binding sites of three different transcription factors with similar binding motifs. Third, we examine differences in motifs of a protein domain.

DNA motifs of same transcription factor

We consider sequence logos and difference logos of binding sites of the human insulator CTCF in different cell lines as obtained by motif discovery from ChIP-seq data [31] based on preprocessed ChIP-seq data from the ENCODE project. For CTCF motif inference, sequences with p-values smaller than 10^{-6} were selected. All data are freely available as Additional File of the original publication [31]. Since CTCF is a DNA-binding protein, the alphabet corresponds to the four nucleotides in this case.

In Fig. 1, we plot the sequence logos for two of these cell types, namely H1-hESC and HUVEC. Considering the sequence logos, both motifs look highly similar with regard to the conservation as well as the nucleotide preference of individual motif positions, and differences between both motifs are hard to perceive. Considering the corresponding difference logo in Fig. 3 (row 1, column 5 or row 5 column 1), however, we instantly see that indeed a large number of motif positions exhibits differences in nucleotide composition. We find the largest difference

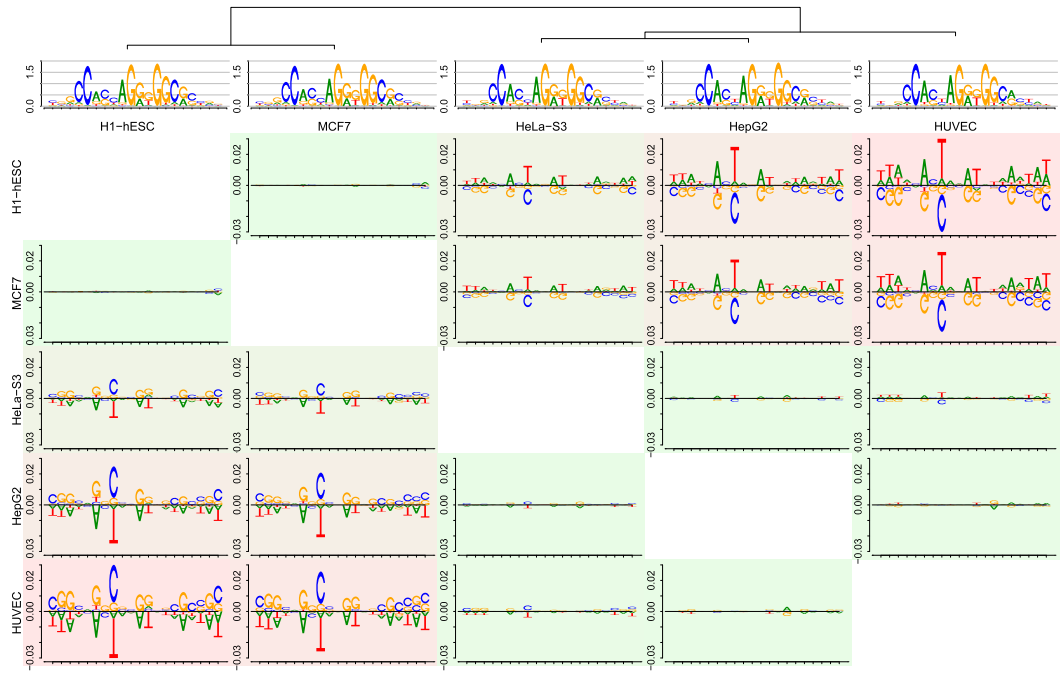


Fig. 3 Comparison of five DNA motifs using *DiffLogo*. Comparison of five CTCF motifs from cell lines H1-hESC, MCF7, HeLa-S3, HepG2, and HUVEC. We plot all pair-wise sequence logos and display the distance between each motif using the background color from green (similar) to red (dissimilar). We plot the sequence logos of each motif as well as the leaf-ordered cluster tree above. The motifs of H1-hESC and MCF7 are highly similar and substantially different from the other motifs, while the motifs of HeLe-S3, HepG2, and HUVEC are similar to each other as well. Due to leaf ordering, the difference between compared motifs increases with increasing distance from the main diagonal in the difference logo grid

according to the difference logo at position 8 of the motifs, where nucleotide C is more prevalent in cell type H1-hESC compared to HUVEC, whereas the opposite holds for nucleotide T. This difference is less visible in the sequence logos, even with hindsight from the difference logo, due to the low conservation at this position. Specifically, the probability of C increases from 0.35 (HUVEC) to 0.58 (H1-hESC), whereas the probability of T drops by a factor of 2 from 0.44 (HUVEC) to 0.21 (H1-hESC). Depending on the application, this difference at position 8 might have a decisive influence on the outcome of, e.g., *in silico* binding site prediction.

In the literature, several positions with substantial motif differences uncovered by *DiffLogo* are known to be related to CTCF binding affinity. For instance [32] show that "low occupancy" CTCF binding sites are enriched for C or G at position 18 compared to "high occupancy" sites, which in our case might indicate that the H1-hESC ChIP-seq data set contains a larger number of such "low occupancy" sites than the HUVEC data set.

In a large-scale study [33], CTCF core motifs are partitioned by the presence or absence of additional upstream and downstream motifs, where the greatest variations in the core motifs between partitions can be found at positions 1-3, 6, 8, 11, 12, 18, and 20, which cover those positions varying in the difference logo. Again, these partitions are related to binding affinity and occupancy of CTCF.

In summary, *DiffLogo* helps to identify several motif positions with substantial variation between cell types, known to be related to CTCF binding affinity and binding site occupancy.

In real-world applications, motifs for more than two cell types are often studied, which might render the pairwise comparison of difference logos a tedious task. We support such an evaluation across multiple cell types by a structured visualization of multiple difference logos as shown in Fig. 3. Here, we compare the pairwise difference logos of CTCF motifs from five cell types, namely H1-hESC, MCF7, HeLa-S3, HepG2, and HUVEC. The cluster tree and background color of the cells are based on numerical measures of motif differences (cf. Implementation) and guide us to the most notable differences between pairs of motifs. For instance, we observe from the tree and background colors that the motifs of H1-hESC and MCF7 are highly similar. The same holds true for the motifs of HeLa-S3, HepG2, and HUVEC, whereas motifs show substantial differences between these two groups. To further facilitate the visual comparison of multiple motifs, we leaf-order the cluster tree such that neighboring motifs are as similar as possible. Due to this ordering, the difference between motif pairs increases with increasing distance from the main diagonal of the difference logo grid. For instance, the topology of the clustering would allow to invert the order of the three leaves under the right sub-tree in Fig. 3, which, however, would bring the quite dissimilar motifs of HUVEC and MCF7 in direct neighborhood. From Fig. 3, we also observe that the two motifs of H1-hESC and HUVEC are the most dissimilar ones among the motifs studied. A visualization of all nine available motifs can be found in Additional file 1: Figure S1.

DNA motifs of different transcription factors

We demonstrate the utility of *DiffLogo* for motifs derived from binding assays for the human transcription factors Max, Myc, and Mad (Mxi1) from Mordelet *et al.* [34]. These three basic helix-loop-helix transcription factors are members of a regulatory network of transcription factors that controls cell proliferation, differentiation, and cell death. Each transcription factor binds to different sets of target sites, regulates different sets of genes, and thus plays a distinct role in human cells. However, Myc, Max, and Mad have almost identical PWMs, which all correspond to an E-box motif with consensus sequence CACGTG.

The PWMs considered here have been derived from probe sequences and corresponding binding intensities of *in-vitro* genomic context protein-binding microarrays [34]. The exact binding sites within the probe sequences are predicted by the de-novo motif discovery tool Dimont [14] using Slim models [35]. For each of the three transcription factors, the top 1,000 predicted binding sites are used to generate the corresponding PWM.

In Fig. 4, we plot the sequence logos and difference logos of Myc, Max, and Mad. We observe from the sequence logos that the binding motifs are almost identical. Considering the difference logos, we observe that the six core nucleotides are conserved in the motifs of all three transcription factors. We find the largest differences between the motif of Max and the motifs of Myc and Mad. In case of Max and Myc, we find a Jensen-Shannon divergence greater than 0.01 bit at positions 11, 12, 22, and 26. In case of Max and Mad, we find a Jensen-Shannon divergence greater than 0.01 bit at positions 3, 12, 22, and 25. In both cases, we mainly find more purine (adenine and guanine) in the motif of Max than in the motifs of Myc and Mad.

Protein motifs

As a third example, we demonstrate the utility of *DiffLogo* using the F-box domain, which plays a role in protein-protein binding. The complete F-box domain in this example is 48 amino acids long [36]. Here, we investigate the middle section from the 12th to the 35th amino acid.

In Fig. 5, we plot the sequence logos and difference logos of F-box domains from the three kingdoms metazoa, fungi, and viridiplantae. We observe from the cluster

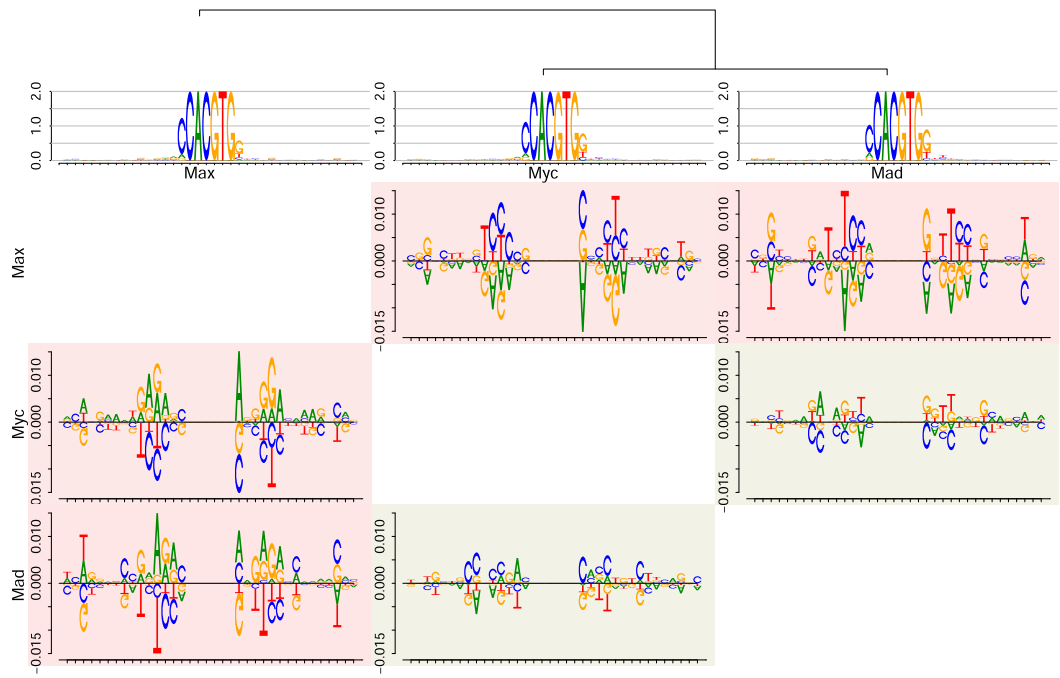


Fig. 4 Comparison of E-Box motifs of Max, Myc, and Mad using *DiffLogo*. We plot all pair-wise difference logos and display the distance between each motif using the background color from green (similar) to red (dissimilar). We plot the sequence logos of each motif as well as the leaf-ordered cluster tree above. The motifs of the transcription factors Myc and Mad are more similar to each other than to the motif Max. The six core nucleotides with consensus sequence CACGTG are conserved in the motifs of all three transcription factors and, hence, are not visible in the difference logos

tree and the background colors that the motifs of metazoa and fungi are highly similar, whereas motifs of this group show substantial differences to viridiplantae. The largest difference can be seen between motifs of metazoa and viridiplantae.

When comparing metazoa and fungi with viridiplantae, *DiffLogo* identifies positions 6, 17, and 22 with high values of the Jensen-Shannon divergence. The differences at positions 6 and 22 could be expected from the differences of the sequence logos, whereas the differences at position 17 are not immediately obvious from them. At position 6 the abundance of arginine (R) in viridiplantae is 0.54 and thus more than 10 times higher than in fungi and 12 times higher than in metazoa. At position 22 tryptophane (W) is highly abundant in viridiplantae and 4 and 3.4 times more abundant than in metazoa and fungi. At position 17 the most noticeable differences in viridiplantae to fungi and metazoa can be seen for amino acid cysteine (C), valine (V), alanine (A), and serine (S). The overall abundance increases from 0.13 in metazoa and 0.12 in fungi to 0.64 in viridiplantae. In contrast, the abundance of arginine (R), glutamine (Q), and lysine (K) is only 0.044 in viridiplantae and 0.44 in metazoa and fungi. A visualization of the full F-Box domain from four kingdoms can be found in Additional file 1: Figure S2.

Conclusion

We present *DiffLogo*, an easy-to-use tool for a fast and efficient comparison of motifs. *DiffLogo* may be applied by users with only basic knowledge in R and is highly configurable and extensible for advanced users. We introduce weighted differences of probabilities to emphasize large differences in position-specific symbol distributions. We present visual comparisons of multiple motifs stemming from motifs of one transcription factor in different cell types, different transcription factors with similar binding motifs, and species-specific protein domains. Figures generated by *DiffLogo* enable the identification of overall motif groups and of sources of dissimilarity. Using *DiffLogo*, it is easily possible to compare motifs from different sources, so *DiffLogo* facilitates decision making, knowledge sharing, and the presentation of results. We make *DiffLogo* freely available in an extensible, ready-to-use R package including examples and documentation. *DiffLogo* is part of *Bioconductor*.

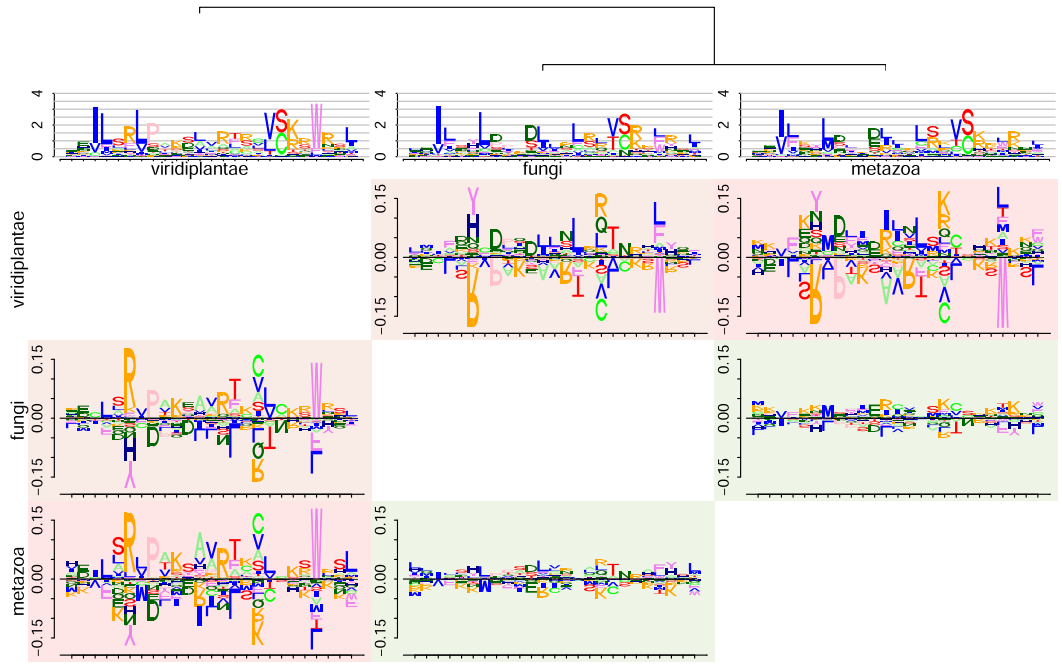


Fig. 5 Comparison of three F-box domain motifs using *DiffLogo*. We compare the F-box domains from the kingdoms metazoa, fungi, and viridiplantae and plot all pair-wise difference logos and display the distance between each motif using the background color from green (similar) to red (dissimilar). We plot the sequence logos of each motif as well as the leaf-ordered cluster tree above. The motifs of metazoa and fungi are highly similar. All other pairwise comparisons show substantial differences

Availability and requirements

Project name: DiffLogo
Project home page: http://github.com/mgledi/DiffLogo
Availability: http://bioconductor.org/packages/DiffLogo
Operating system(s): Platform independent
Programming language: R
Other requirements: Installation of R 1.8.0 or higher
License: LGPL ($\geq$ 2)
Any restrictions to use by non-academics: None

Additional file

Additional file 1: Supplementary Methods, Results, Figures, and Examples. This file is structured in four sections. Section 1, *Additional examples*, contains Figures S1 and S2. Figure S1 shows a *DiffLogo* grid for nine CTCF motifs. Figure S2 shows a *DiffLogo* grid for four F-box domain motifs. In section 2, *CTCF with and without clustering*, we show in detail the impact of clustering and optimal leaf ordering for a *DiffLogo* grid of nine CTCF motifs. In section 3, *Alternative combinations of stack heights and symbol weights*, we first describe the mathematical background of four implementations of H_ℓ and two implementations of $r_{\ell,a}$. Afterwards, we show the result of the eight possible combinations in Tables S1 and S2 on two sequence motifs. In section 4, *Tool comparison*, we compare *DiffLogo* with the five tools *seqLogo*, *iceLogo*, *MotifStack*, *STAMP*, and *Two Sample Logo*.

From the set of nine CTCF motifs we selected the pair of motifs with the highest similarity according to the Jensen-Shannon divergence (GM12878 and K562) and the pair of motifs with the lowest similarity according to the Jensen-Shannon divergence (H1-hESC and HUVEC) for the comparison of the five different tools. (PDF 8775 kb)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MN conceived the idea. MN, HT, JK, JG, SP, and IG developed the idea and the computational methods. MN and HT implemented and tested *DiffLogo*. All of the authors read and approved the final version of the manuscript.

Acknowledgements
We thank Karin Breunig, Jesus Cerquides, Ralf Eggeling, and Martin Porsch for valuable discussions and contributing data and DFG (grant no. GR3526/1) for financial support.

Author details
[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany. [2]Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany. [3]Institute for Biosafety in Plant Biotechnology, Julius Kühn-Institut (JKI), Federal Research Centre for Cultivated Plants, Quedlinburg, Germany. [4]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany.

Received: 10 April 2015 Accepted: 8 October 2015
Published online: 17 November 2015

References

1. Staden R. Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. 1984;12:505–19.
2. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. San Diego: Department of Computer Science and Engineering, University of California; 1994.
3. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268(1):78–94.
4. Yeo G, Burge CB. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. J Comput Biol. 2004;11(2–3):377–94.
5. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, et al. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 2010;38(Database issue):161–6. doi:10.1093/nar/gkp885.
6. Elnitski L, Jin VX, Farnham PJ, Jones SJM. Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. Genome Res. 2006;16:4140006.
7. Zhao Y, Granas D, Stormo GD. Inferring binding energies from selected binding sites. PLoS Comput Biol. 2009;5(12):1000590.
8. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, et al. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. Genome Res. 2010;20(6):861–73.
9. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316(5830):1497–502.
10. Galas DJ, Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 1978;5(9):3157–170. doi:10.1093/nar/5.9.3157.
11. Bailey TL, Williams N, Misleh C, Li WW. Meme: discovering and analyzing dna and protein sequence motifs. Nucleic Acids Res. 2006;34(Web-Server-Issue):369–73.
12. Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ. Deep and wide digging for binding motifs in chip-seq data. Bioinforma. 2010;26(20):2622–23.
13. Ma X, Kulkarni A, Zhang Z, Xuan Z, Serfling R, Zhang MQ. A highly efficient and effective motif discovery method for chip-seq/chip-chip data using positional information. Nucleic Acids Res. 2012;40(7):50.
14. Grau J, Posch S, Grosse I, Keilwagen J. A general approach for discriminative de novo motif discovery from high-throughput data. Nucleic Acids Res. 2013;41(21):197.
15. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18(20):6097–100.
16. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004;32(Database issue):91–4.
17. Newburger DE, Bulyk ML. Uniprobe: an online database of protein binding microarray data on protein–dna interactions. Nucleic Acids Res. 2009;37(suppl 1):77–82.
18. Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae1. J Mol Biol. 2000;296(5):1205–14. doi:10.1006/jmbi.2000.3519.
19. Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B. Computational detection of cis -regulatory modules. Bioinformatics. 2003;19(suppl 2):5–14. doi:10.1093/bioinformatics/btg1052.
20. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004;431(7004):99–104. doi:10.1038/nature02800.
21. Linhart C, Halperin Y, Shamir R. Transcription factor and microrna motif discovery: The amadeus platform and a compendium of metazoan target sets. Genome Research. 2008;18(7):1180–9. doi:10.1101/gr.076117.108.
22. Bembom O. SeqLogo: Sequence logos for DNA sequence alignments. 2015. http://www.bioconductor.org/packages/release/bioc/html/seqLogo.html, accessed 2015.03.05.
23. Colaert N, Helsens K, Martens L, Vandekerckhove J, Gevaert K. Improved visualization of protein consensus sequences by iceLogo. Nat Meth. 2009;6(11):786–7. doi:10.1038/nmeth1109-786.
24. Jianhong Ou LJZ. MotifStack: Plot Stacked Logos for Single or Multiple DNA, RNA and Amino Acid sequence. http://www.bioconductor.org/packages/release/bioc/html/motifStack.html. Accessed on 13 Feb 2015.
25. Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. 2007;35(Web Server issue):272–58. doi:10.1093/nar/gkm272.
26. Vacic V, Iakoucheva LM, Radivojac P. Two sample logo: a graphical representation of the differences between two sets of sequence alignments. Bioinforma. 2006;22(12):1536–7. doi:10.1093/bioinformatics/btl151.
27. Ali SM, Silvey SD. A general class of coefficients of divergence of one distribution from another. J R Stat Soc Series B (Methodological). 1966;28(1):131–42.
28. Lin J. Divergence measures based on the Shannon entropy. Inf Theory, IEEE Trans on. 1991;37(1):145–51. doi:10.1109/18.61115.
29. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. http://www.R-project.org/.
30. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome biology. 2004;5(10):80–16. doi:10.1186/gb-2004-5-10-r80.
31. Eggeling R, Gohr A, Keilwagen J, Mohr M, Posch S, Smith AD, et al. On the value of intra-motif dependencies of human insulator protein ctcf. PLoS ONE. 2014;9(1):85629. doi:10.1371/journal.pone.0085629.
32. Plasschaert RN, Vigneau S, Tempera I, Gupta R, Maksimoska J, Everett L, et al. CTCF binding site sequence differences are associated with unique regulatory and functional trends during embryonic stem cell differentiation. Nucleic acids research. 2014;42(2):774–89. doi:10.1093/nar/gkt910.
33. Nakahashi H, Kwon K-RKR, Resch W, Vian L, Dose M, Stavreva D, et al. A genome-wide map of CTCF multivalency redefines the CTCF code. Cell reports. 2013;3(5):1678–89. doi:10.1016/j.celrep.2013.04.024.
34. Mordelet F, Horton J, Hartemink AJ, Engelhardt BE, Gordân R. Stability selection for regression-based models of transcription factor-DNA binding specificity. Bioinforma. 2013;29(13):117–25. doi:10.1093/bioinformatics/btt221.
35. Keilwagen J, Grau J. Varying levels of complexity in transcription factor binding motifs. Nucleic Acids Res. 2015;43(18):e119.
36. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(D1):222–30. doi:10.1093/nar/gkt1223.

7 Appendix

The following sections contain important additional studies important for the understanding of this thesis. More supplementary studies, figures, and tables can be found in the additional files of the corresponding articles.

7.1 Detecting and correcting the binding-affinity bias in ChIP-Seq data using inter-species information

The supplementary material of "Detecting and correcting the binding-affinity bias in ChIP-seq data using inter-species information" consists of two additional files. **Additional File 1** consists of 3 sections. In **Section 1**, *Modeling the binding-affinity bias*, we describe how to determine the likelihood of non-motif-bearing and motif-bearing alignments modeling the contamination bias and the binding-affinity bias. In **Section 2**, *Example interpretation of difference logos*, we give an exemplary interpretation of some difference logos. *Section 3*, *Supplementary Figures*, contains **Supplementary Figures S1-S18**. **Additional File 2** contains the sequence data used in the studies of this work. Here, I provide a copy of **Section 1** of **Additional File 1**. This section is the mathematical counter part to the section "Modeling the binding-affinity bias" in "Methods" in the main manuscript, where modeling the binding-affinity bias is explained from the data generating perspective.

7.1.1 Modeling the binding-affinity bias

In this section we describe the probabilistic model for modeling the binding-affinity bias. We define the model in mathematical terms by providing the likelihood function. We use the notation from the manuscript.

Following the data-generating process described in the manuscript, the probability that the model generates an alignment X_n can be written as

$$\begin{aligned} p(X_n|\theta) &= p(X_n|M_n = 0, \theta) \cdot p(M_n = 0, \theta) + p(X_n|M_n = 1, \theta) \cdot p(M_n = 1, \theta) \\ &= p(X_n|M_n = 0, \theta) \cdot \alpha + p(X_n|M_n = 1, \theta) \cdot (1 - \alpha) \end{aligned}$$

To complete the model, we need to specify the probability for non-motif-bearing alignments $p(X_n|M_n = 0, \theta)$ and that for motif-bearing alignments $p(X_n|M_n = 1, \theta)$.

Likelihood of a non-motif-bearing alignment

Looking at the description of the generating process for non-motif-bearing alignments we get

$$p(X_n|M_n = 0, \theta) = \sum_{Y_n \in \mathcal{A}^{L_n}} p(Y_n|M_n = 0, \theta) \prod_{o=1}^{O} p(X_n^{\cdot,o}|Y_n, M_n = 0, \theta).$$

Note that given θ and $M_n = 0$, each single nucleotide alignment is independent of any other single nucleotide alignment. Thus, the likelihood can be expressed as

$$p(X_n|M_n = 0, \theta) = \prod_{u=1}^{L_n} \sum_{Y_n^u \in \mathcal{A}} p(Y_n^u|M_n = 0, \theta) \prod_{o=1}^{O} p(X_n^{u,o}|Y_n^u, M_n = 0, \theta).$$

Here we denote $p(Y_n^u|M_n = 0, \theta)$ and $p(X_n^{u,o}|Y_n^u, M_n = 0, \theta)$ by parameters

$$\begin{aligned} p(Y_n^u|M_n = 0, \theta) &= \pi_0^{Y_n^u} \\ p(X_n^{u,o}|Y_n^u, M_n = 0, \theta) &= \gamma_o \cdot \pi_0^{X_n^{u,o}} + (1 - \gamma_o) \cdot \delta_{X_n^{u,o} = Y_n^u} \end{aligned}$$

according to the F81 model, where the base distribution of each position of the background sequence is denoted by π_0, the probability of a nucleotide a in the background sequence is denoted by π_0^a, and the substitution probability from the primordial species to species o is denoted by γ_o.

Likelihood of a motif-bearing alignment

In the data generating process for motif-bearing alignments we sample alignments until one of them is accepted. Mapping this into a likelihood requires the usage of the *Felsenstein's pulley principle* Felsenstein, 1981, that allows us to select any particular species as the root of the tree. In this case it will come handy to select the reference species as the root. Thus, the likelihood can be expressed as

$$\begin{aligned} p(X_n|M_n = 1, \theta) = \sum_{\ell_n=1}^{L_n - W + 1} p(X_n^{\cdot,1}|M_n = 1, \ell_n) \cdot \sum_{Y_n \in \mathcal{A}^{L_n}} p(Y_n|X_n^{\cdot,1}, M_n = 1, \ell_n) \cdot \\ \prod_{o=2}^{O} p(X_n^{\cdot,o}|Y_n, M_n = 1, \ell_n) p(\ell_n|M_n = 1, \theta), \end{aligned}$$

where the base distributions of the positions $1, \ldots, W$ of the binding sites are denoted by $\pi_1, \ldots, \pi_W$ and the probability of a nucleotide a in the binding site at position w is denoted by π_w^a.

Given π, $\ell_n \in \{1, \ldots, L_n - W + 1\}$, and $M_n = 1$, each single nucleotide alignment is independent of any other single nucleotide alignment, and we obtain

$$p(X_n|M_n = 1, \theta) = \sum_{\ell_n=1}^{L_n-W+1} \prod_{u=1}^{L_n} p(X_n^{u,1}|M_n = 1, \ell_n) \cdot \sum_{Y_n^u \in \mathcal{A}} p(Y_n|X_n^{u,1}, M_n = 1, \ell_n) \cdot \prod_{o=2}^{O} p(X_n^{u,o}|Y_n, M_n = 1, \ell_n) p(\ell_n|M_n = 1, \theta).$$

We need to determine the probability of a particular nucleotide in a specific position of the reference species after selection, that is $p(X_n^{u,1}|M_n = 1, \ell_n)$. On one hand, notice that selection does not affect the probability distribution of those nucleotides outside the binding site. Thus, for $u < \ell_n$ or $u \geq \ell_n + W$ we have that $p(X_n^{u,1} = a|M_n = 1, \ell_n) = \pi_0^a$. On the other hand, for nucleotides in the binding site, the distribution after filtering is $p(X_n^{u,1} = a|M_n = 1, \ell_n) \propto (\pi_{u-\ell_n+1}^a)^\beta$. Thus, $p(X_n^{u,1} = a|M_n = 1, \ell_n) = \frac{(\pi_{u-\ell_n+1}^a)^\beta}{\sum_{b \in \mathcal{A}} (\pi_{u-\ell_n+1}^b)^\beta}$.

The probabilities for the nucleotides in the ancestral sequence and in the non-reference species are given by the F81 model. In particular, for the ancestral sequence

$$p(Y_n = a|X_n^{u,1} = b, M_n = 1, \ell_n) = \begin{cases} \gamma_1 \cdot \pi_0^a + (1 - \gamma_1) \cdot \delta_{a=b} & \text{, if } u < \ell_n \text{ or } u \geq \ell_n + W \\ \gamma_1 \cdot \pi_{u-\ell_n+1}^a + (1 - \gamma_1) \cdot \delta_{a=b} & \text{, if } \ell_n \leq u < \ell_n + W \end{cases}$$

and for the non reference species

$$p(X_n^{u,o} = a|Y_n = b, M_n = 1, \ell_n) = \begin{cases} \gamma_o \cdot \pi_0^a + (1 - \gamma_o) \cdot \delta_{a=b} & \text{, if } u < \ell_n \text{ or } u \geq \ell_n + W \\ \gamma_o \cdot \pi_{u-\ell_n+1}^a + (1 - \gamma_o) \cdot \delta_{a=b} & \text{, if } \ell_n \leq u < \ell_n + W \end{cases}$$

Finally, since we assume binding sites to be uniformly distributed, we have that $p(\ell_n|M_n = 1, \theta) = \frac{1}{L_n - W + 1}$. This completes the specification of the likelihood function.

7.2 Unrealistic phylogenetic trees may improve *phylogenetic footprinting*

The supplementary material of "Unrealistic phylogenetic trees may improve phylogenetic footprinting" consists of one additional file that contains Supplementary Methods, Results, Figures, and Examples. This file comprises five sections. In **Section 1**, *Accuracy of predicted motifs*, we scrutinize the motifs obtained by PFMs with a substitution probability of $\gamma = 1.0$. **Section 2**, *Likelihood, classification performance, and difference logos for 5 transcription factors*, contains supplementary Figures to the studies on the five TFs presented in the main manuscript. In **Section 3**, *Comparison of classification performances of PFMs basing on five different phylogenetic trees*, we extend the study presented in the main manuscript and compare the calssification performance on the five PFMs $\mathcal{M}_{lit}^{tree}$, $\mathcal{M}_{ML}^{star}$, $\mathcal{M}_{\gamma=1.0}^{star}$, $\mathcal{M}_{\hat{\gamma}}^{tree}$, and $\mathcal{M}_{\hat{\gamma}}^{star}$. In **Section 4**, *Synthetic tests*, we provide exemplary studies on synthetic data. **Section 5**, *Supplementary Tables*, comprises tables regarding related phylogenetic footprinting approaches, dataset statistics, and P-values for the in the main manuscript presented results. Here, I provide a copy of **Section 1** and **Section 3**.

7.2.1 Accuracy of predicted motifs

In the main manuscript, we show that on real data PFMs basing on unrealistic substitution probabilities (unrealistic PFMs) outperform PFMs basing on realistic substitution probabilities (realistic PFMs) in contrast to synthetic data where realistic PFMs outperform unrealistic PFMs. Here, we investigate the degree of similarity between the motifs inferred with realistic PFMs and unrealistic PFMs in two studies. First, on synthetic data, we compare the accuracy of inferred motifs for different combinations of substitution probabilities used for data generation and for motif inference. Second, on real data, we compare the motif similarity of the motif inferred using an unrealistic PFM to the motifs inferred using more realistic PFMs.

7.2.1.1 Test on synthetic data

We study on synthetic data to which amount different substitution probabilities for data generation and different substitution probabilities for the inference of a PFM affect the accuracy of *de–novo* motif prediction. We generate synthetic datasets basing on different substitution probabilities and we infer on each synthetic dataset a set of PWMs using PFMs basing different substitution probabilities as follows.

First, we generate for each substitution probability $\alpha = \{0.1, 0.2, \ldots, 1.0\}$ a dataset consisting of $N = 1000$ motif alignments of length $W = 10$ each with $O = 5$ species. The set of

ancestor sequences is sampled from a PWM π of length W whose probability distribution is generated randomly. Each ancestor sequence is mutated using the F81 model with a star topology with all O substitution probabilities set to α.

Second, for each generated dataset we estimate for each substitution probability $\gamma = \{0.1, 0.2, \ldots, 1.0\}$ a PFM with a star topology with all substitution probabilities set to γ. For each estimated PFM we extract the PWM $\hat{\pi}_\gamma$ and quantify the dissimilarity between $\hat{\pi}_\gamma$ and π by the symmetric Kullback—Leibler divergence (KLD). A KLD equal to 0 indicates identical PWMs $\hat{\pi}_\gamma$ and π. The KLD is proportional to the degree of dissimilarity between the PWMs $\hat{\pi}_\gamma$ and π.

We repeat both steps 50 times and determine the mean KLD for each combination of $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$ and $\gamma \in \{0.1, 0.2, \ldots, 1.0\}$. In **Figure 7.1** we show the mean KLD for each combination of α and γ.

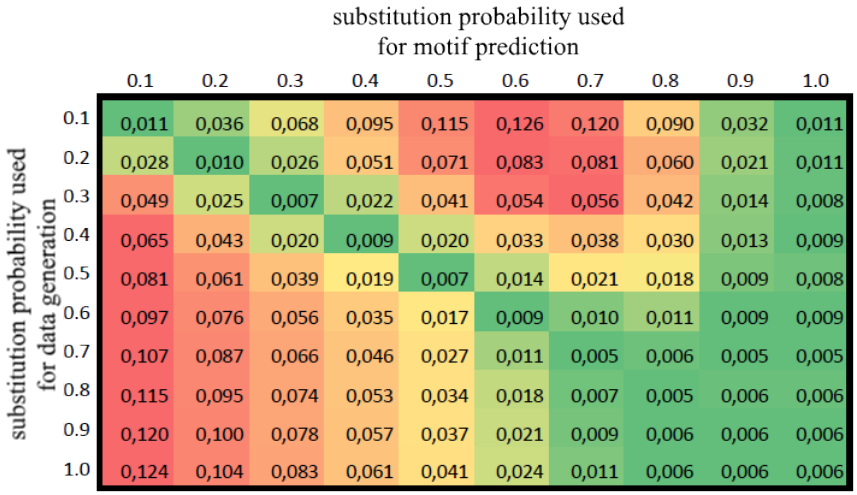


Figure 7.1: Motif accuracy for different combinations of substitution probabilities used for data generation and substitution probabilities used for motif inference. We represent the datasets generated with a star topology with substitution probabilities set to $\alpha = \{0.1, 0.2, \ldots, 1.0\}$ in the rows. We represent the PFMs basing on substitution probabilities $\gamma = \{0.1, 0.2, \ldots, 1.0\}$ in the columns. For each combination of α and γ, we specify the mean KLD of the true and the estimated motif and we visualize these values with the background color from green (similar) to red (dissimilar). For each row, we find the smallest KLD for $\alpha = \gamma$ and we find highly similar results for $\gamma = 1.0$.

For each dataset we find minimal KLDs for the PWM $\hat{\pi}_{\gamma=\alpha}$ and the PWM $\hat{\pi}_{\gamma=1.0}$. With other words, the motifs inferred with a PFM basing on a substitution probabilities with γ equal to α (the substitution probability that was used for data generation) and the motifs inferred with a PFM basing on substitution probabilities γ equal to 1.0 (which implies conditional independence between sequences) are highly similar. The KLDs of the PWMs

$\hat{\pi}_\gamma$ with $\gamma \neq \alpha$ and $\gamma \neq 1.0$ are greater than or equal to the KLDs for the PWMs $\hat{\pi}_{\gamma=\alpha}$ and $\hat{\pi}_{\gamma=1.0}$ in every case.

7.2.1.2 Test on real data

In the previous study on synthetic data we have shown that the PWM inferred using the most realistic PFM ($\hat{\pi}_{\gamma=\alpha}$) and the PWM inferred using the most unrealistic PFM ($\hat{\pi}_{\gamma=1.0}$) are most similar. We study whether this relationship is also true in case of real data. In case of real data we do not know the true α, i.e., the substitution probability between the ancestor species and the observed species–specific sequences. Hence, we compare the PWM $\hat{\pi}_{1.0}$ with the PWMs $\hat{\pi}_\gamma$ for $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$. Again, we quantify the dissimilarity between $\hat{\pi}_{1.0}$ and $\hat{\pi}_\gamma$ by the symmetric KLD.

For each of the five TFs described in **Methods 1** and each decomposition of the 100–fold stratified repeated random sub–sampling validation procedures described in **Methods 4**, we calculate the KLDs of the PWM $\hat{\pi}_{1.0}$ and the PWMs $\hat{\pi}_\gamma$ for $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$ inferred on the positive training dataset. We compute mean and standard error of the resulting 100 KLDs for each pair of the the PWM $\hat{\pi}_{1.0}$ and the PWMs $\hat{\pi}_\gamma$. We show mean and standard error of the KLDs as function of γ in **Figure 7.2** for each of the five TFs. Based on the previous study and the results presented by Gertz *et. al* 2006 (Gertz et al., 2006), we expect a local minimum of the KLDs for $0.1 < \gamma \leq 0.4$.

We find for each TF a local minimum of the KLD between the PWM $\hat{\pi}_{1.0}$ and the PWM $\hat{\pi}_\gamma$ for realistic substitution probabilities $0.1 < \gamma \leq 0.4$ ($\gamma = 0.35$ for CTCF, $\gamma = 0.3$ for GABP, $\gamma = 0.25$ for NRSF, $\gamma = 0.4$ for SRF, $\gamma = 0.35$ for STAT1). For γ smaller than these minimums the KLD increases monotonically and for γ greater than these minimums the KLD first increases, reaches a local maximum for $0.6 \leq \gamma \leq 0.7$, and again decreases for γ greater this local maximum (with a KLD equal to zero for $\gamma = 1.0$ per definition). In accordance to the results on synthetic data, we show that the motifs inferred using a PFM basing on unrealistic substitution probabilities are similar to the motifs inferred using a PFM basing on realistic substitution probabilities. Since the true substitution probabilities are not known in case of real data, the estimation of motifs using an unrealistic PFM is a potential more robust way to avoid errors from falsely estimated substitution probabilities, i.e., unrealistic PFMs seem to be more robust against model violations in the dataset.

7.2.2 Synthetic tests

In the main manuscript, we show that PFMs with unrealistic substitution probabilities outperform realistic PFMs on real data in contrast to synthetic data. Here, we investigate

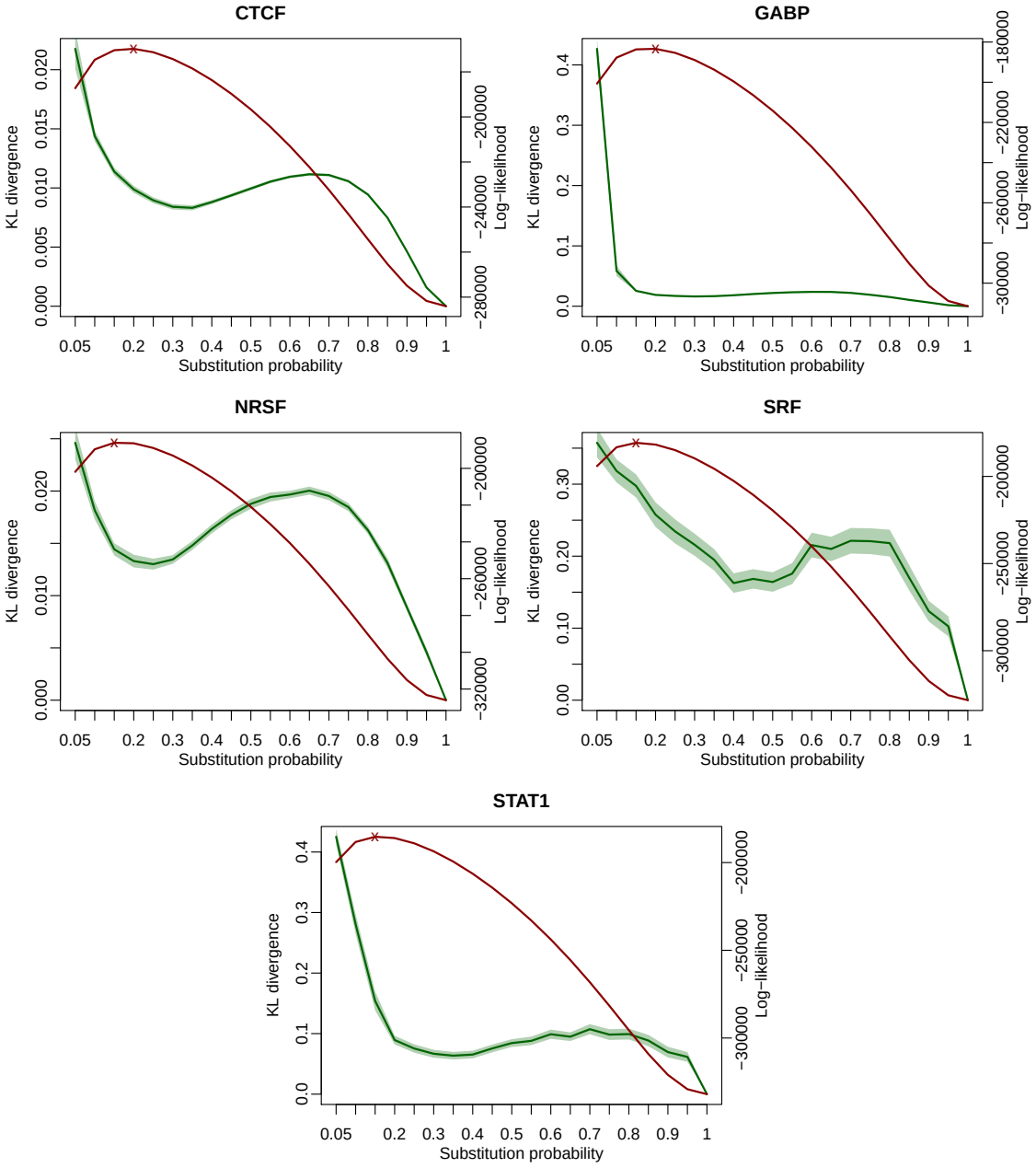


Figure 7.2: Motif dissimilarity between $\hat{\pi}_{1.0}$ and $\hat{\pi}_{\gamma}$ for CTCF, GABP, NRSF, SRF, and STAT1. For each TF, we plot the mean and standard error of the KLD between the PWM $\hat{\pi}_{1.0}$ and the PWMs $\hat{\pi}_{\gamma}$ with $\gamma = \{0.05, 0.1, \ldots, 1.0\}$ (green line). In addition, we plot the mean and standard error of the likelihood of the corresponding PFMs (red line). We find for the KLD a local minimum for $0.1 < \gamma \leq 0.4$ in every case.We find for the likelihoods a global maximum for $0.1 < \gamma \leq 0.4$ in every case (red cross).

the influence of various data properties on classification performance in order to reproduce this observation on synthetic data.

We generate synthetic datasets as described in **Methods 2** and modify this procedure in different ways as follows. We vary the ratio of the size of positive and negative test data in **section 7.2.2.1**, we use different trees for data generation instead of a star in **section 7.2.2.2**, we model heterogeneity during data generation in **section 7.2.2.3**, we use the more realistic HKY evolutionary model instead of the F81 model for data generation in **section 7.2.2.4**, and we use different trees in combination with the more realistic HKY evolutionary model for data generation in **section 7.2.2.5**. All datasets are available at `https://github.com/mgledi/PhyFoo/tree/master/data/synthetic_data/`.

We apply the PFMs described in **Methods 1** on each of the generated datasets with varying substitution probability γ of the PFMs from 0.05 to 1.0 with increments of 0.05 as described in **Methods 3**. We study the classification performance of the PFMs by the method described in **Methods 4**.

7.2.2.1 Unbalanced positive and negative test data

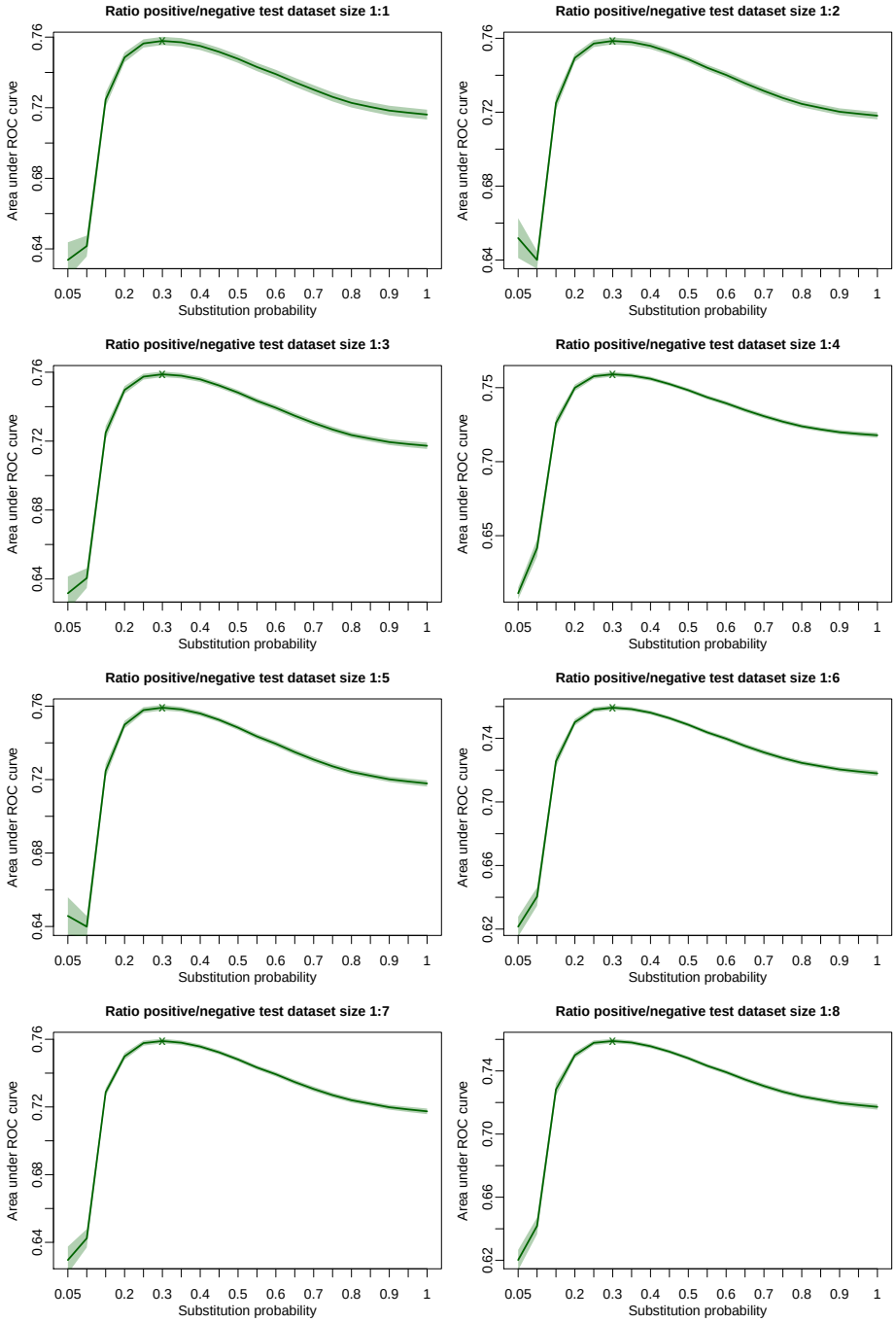


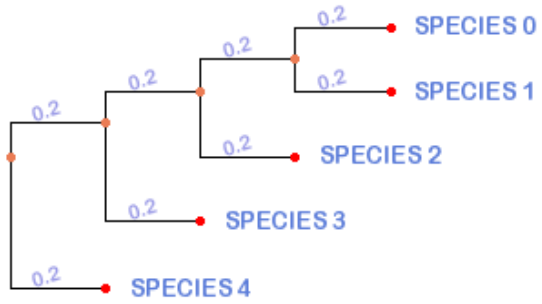
Figure 7.2: Classification performance for different substitution probabilities on synthetic data. We plot classification performance on synthetic data for a PFM using a star topology with all substitution probabilities set to $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$, where the ratio between positive and negative test data is chosen as $1:1$, $1:2$, $1:3$, $1:4$, $1:5$, $1:6$, $1:7$, and $1:8$ respectively. The classification performance behaves as expected.

7.2.2.2 Using trees for data generation

Tested data generation with the following three trees with five species each and all branches having the length $\gamma = 0.2$. Find below the Newick representation of the trees and the corresponding visualisation Fredslund, 2006.

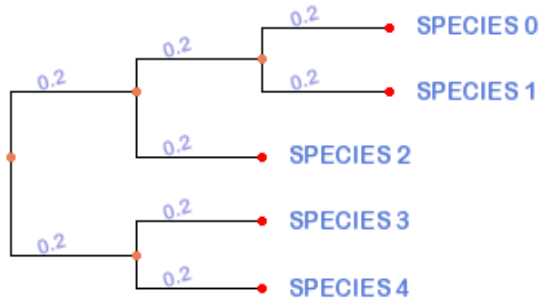
Unbalanced binary tree:

```
(
  (
    (
      (
        SPECIES_0:0.2,
        SPECIES_1:0.2
      ):0.2,
      SPECIES_2:0.2
    ):0.2,
    SPECIES_3:0.2
  ):0.2,
  SPECIES_4:0.2
)
```



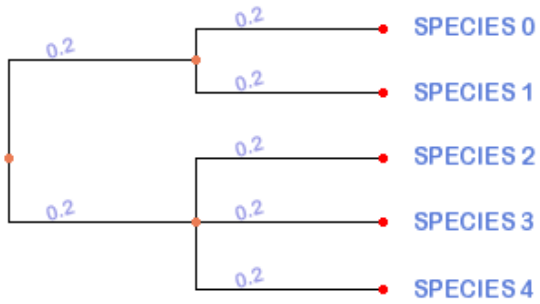
Balanced binary tree:

```
(
  (
    (
      SPECIES_0:0.2,
      SPECIES_1:0.2
    ):0.2,
    SPECIES_2:0.2
  ):0.2,
  (
    SPECIES_3:0.2,
    SPECIES_4:0.2
  ):0.2
)
```



Balanced ternary tree:

```
(
   (
      SPECIES_0:0.2,
      SPECIES_1:0.2
   ):0.2,
   (
      SPECIES_2:0.2,
      SPECIES_3:0.2,
      SPECIES_4:0.2
   ):0.2
)
```



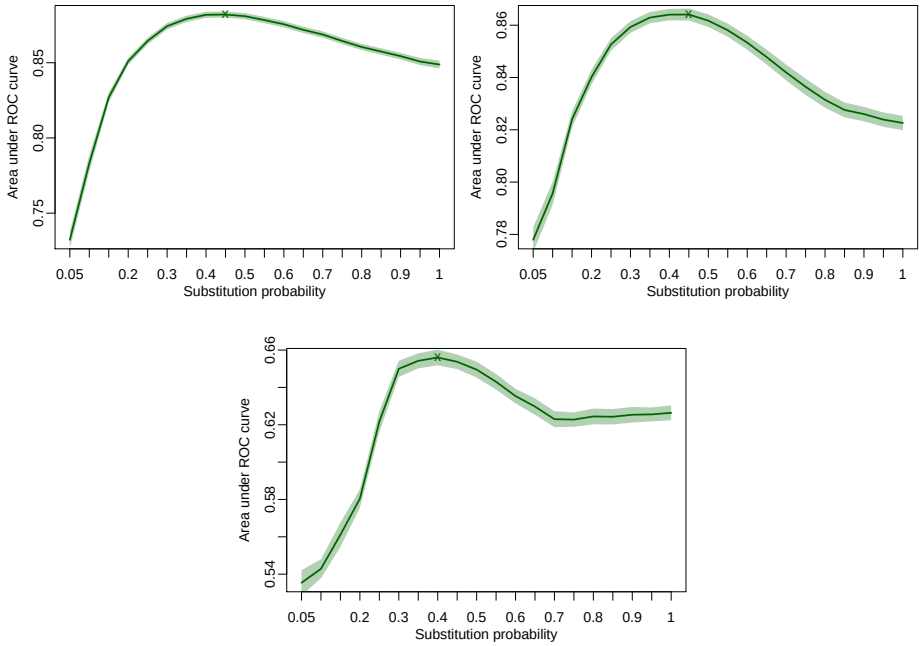


Figure 7.3: Classification performance for different substitution probabilities on synthetic data. We plot classification performance on synthetic data for a PFM using a star topology with all substitution probabilities set to $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$, where the data was generated using (i) an unbalanced binary tree, (ii) a balanced binary tree, and (iii) a balanced ternary tree respectively (trees shown above). The classification performance behaves as expected.

7.2.2.3 Heterogeneity

Tested data generation with three different combinations of heterogeneity. In case of enabled heterogeneity for motif generation each position in each binding-site is generated using an individual star topology with each substitution probability drawn individually from $beta(3, 10)$. In case of enabled heterogeneity for background and flanking region every position in the alignments that does not correspond to a binding site is generated using an individual star topology with each substitution probability drawn individually from $beta(3, 10)$.

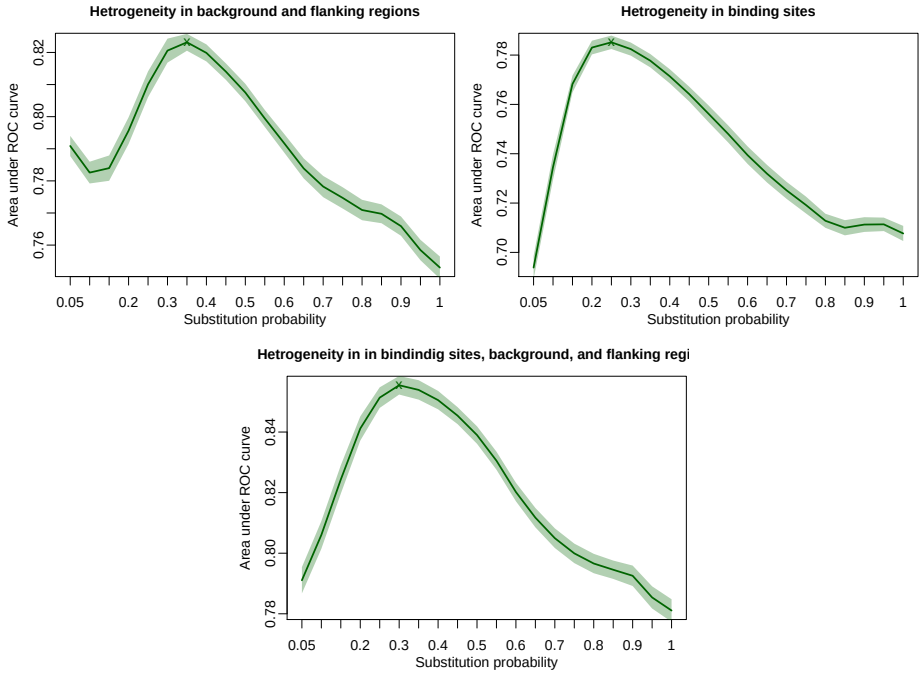


Figure 7.4: Classification performance for different substitution probabilities on synthetic data. We plot classification performance on synthetic data for a PFM using a star topology with all substitution probabilities set to $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$, where the data was generated using heterogenous substitution probabilities (i) only in the background and flanking regions, (ii) only in the binding sites, and (iii) in the background, the flanking regions, and the binding sites respectively. The classification performance behaves as expected.

7.2.2.4 Using HKY model for data generation

Tested data generation using HKY model with two different combinations of transversion and transition probability.

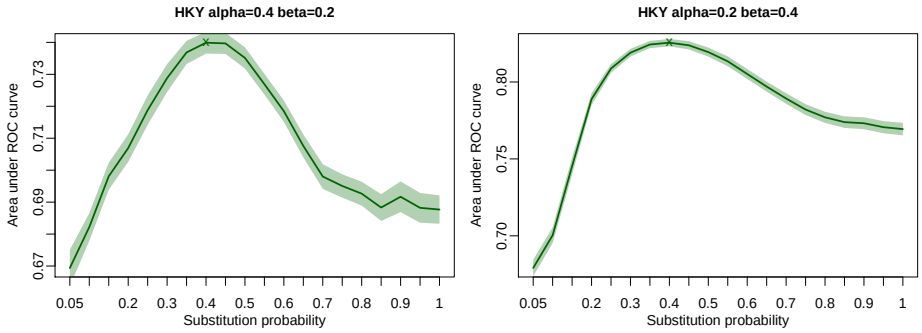


Figure 7.5: Classification performance for different substitution probabilities on synthetic data. We plot classification performance on synthetic data for a PFM using a star topology with all substitution probabilities set to $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$, where the data was generated using the HKY evolutionary model with (i) $\alpha = 0.4$ and $\beta = 0.2$ and (ii) $\alpha = 0.2$ and $\beta = 0.4$ respectively. The classification performance behaves as expected.

7.2.2.5 Using more complex phylogenetic trees with HKY model for data generation

Tested data generation using three different trees with the HKY model with two different combinations of transversion and transition probability.

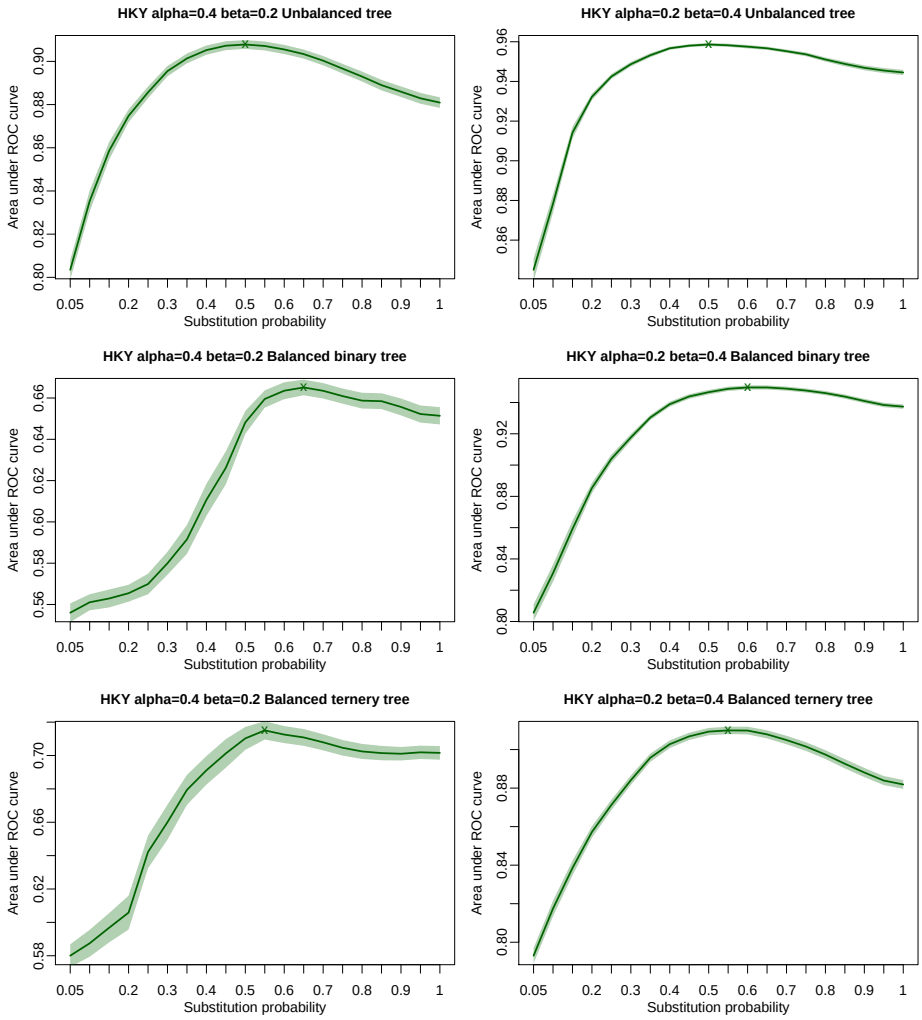


Figure 7.6: Classification performance for different substitution probabilities on synthetic data. We plot classification performance on synthetic data for a PFM using a star topology with all substitution probabilities set to $\gamma \in \{0.05, 0.1, \ldots, 1.0\}$, where the data was generated using the HKY evolutionary model with (Ai) $\alpha = 0.4$ and $\beta = 0.2$ and (Aii) $\alpha = 0.2$ and $\beta = 0.4$ in combination with (Bi) an unbalanced binary tree, (Bii) a balanced binary tree, and (Biii) a balanced ternary tree respectively (trees shown above). The classification performance behaves as expected.

7.3 Combining *phylogenetic footprinting* with motif models incorporating intra-motif dependencies

The supplementary material of "Detecting and correcting the binding-affinity bias in ChIP-seq data using inter-species information" consists of five additional files. **Additional File 1** contains for each of the 35 TFs a 10×10 table of difference logos for a pair-wise visual comparison of species-specific motifs. **Additional File 2** contains for each of the 35 TFs the sequence logo inferred using the PFM(2) aligned with mutual information profiles of order 1, the mutual information profiles of order 2, and species-specific mutual information profiles of orders 1 and 2 for each of the 10 species. **Additional File 3** contains sequence logos and their reverse complements of predicted binding sites inferred using the PFM(0), the PFM(1), and the PFM(2) for each of the 35 TFs. **Additional File 4** contains for each TF two plots showing the 25 ROC curves and the 25 PR curves from the 25–fold stratified repeated random sub-sampling validation procedure described in **Methods 3**. **Additional File 5** This file contains three supplementary sections, presenting four additional studies, details about the implementation and some statistics regarding the datasets of all 35 TFs. **Additional File 6** contains data files of alignments of the ChIP-seq positive regions and negative control regions for each of the 35 TFs in FASTA format.

Here, we show **Section 1.1** and **Section 1.3** of **Additional File 5**. In the first subsection (former section 1.1), we study differences among species–specific motifs of 35 TFs. In the second subsection (former section 1.3), we examine the impact of base dependencies and phylogenetic dependencies on classification performance.

7.3.1 Species–specific motifs are highly similar for most TF

Intra-motif dependencies may be a constant phenomenon conserved across the examined species or a rather dynamic phenomenon significantly changing during the evolution of these species. The latter case may imply that species–specific motifs are different to a certain degree. Consequently, the estimation of base dependencies across species may result in the estimation of spurious results. Hence, **first** we visually study differences among the species–specific motifs for each of the 35 TFs using *difference logos*, **second** we determine whether observable differences between species–specific motifs are significant or not, and **third** we examine the distribution of position-specific MIs for each species. Therefore, we extracted for each of the 35 TFs one motif for each of the 10 species resulting in 35×10 species–specific motifs as described in **Supplementary Section 2.1**. Please note that the extracted species–specific motifs for other species than the reference species are not representative for these phylogenetically related species.

7.3.1.1 Primates show almost no differences in their sequence logos

We use the freely available R package *DiffLogo* for the visual inspection of motif differences Nettling, Treutler, Grau, et al., 2015. *DiffLogo* enables the illustration and investigation of differences between highly similar motifs such as binding motifs of TFs from different experiments, different motif prediction algorithms, or different species. Hence, we use tables of difference logos generated with *DiffLogo* for a pair-wise comparison of all species–specific motifs. Each difference logo displays position-specific differences of base distributions by a stack of bases which height is proportional to the base distribution difference quantified by the Jensen-Shannon divergence. The Jensen-Shannon divergence is zero in case of two identical base distributions and 1 in case of two maximally different base distributions. The tables of difference logos for all 35 TFs can be found in Additional File 1. All sequence logos of PFM(0), PFM(1), and PFM(2) can be found in Additional File 3.

Exemplary, **Supplementary Figure 7.7** shows the table of difference logos for the TF Bach1. We find that the species–specific motifs segregate into two main groups, where one group comprises seven higher primates and the second group comprises three species from the Laurasiatheria superorder, i.e., dog, horse, and cow. We find differences between the motifs of both groups at various motif positions, where the motif differences of relatively high degree are located at rather conserved motif positions as well as at more variable motif positions. For instance, we find relatively high differences at motif position 8, where guanine is more abundant in the primate motifs and the remaining bases are more abundant in the Laurasiatheria motifs. However, the maximum Jensen-Shannon divergence in all difference logos for Bach1 is below 0.01 bits.

We examine the motif differences between species–specific motifs for all 35 TFs. We see for 14 TFs that the set of ten species segregates into the two groups of seven higher primates and three from the Laurasiatheria clade as before in case of Bach1 (CEBPB, CTCF, EGR1, MafK, Max, NRSF, POU5F1, Rad21, SRF, TCF12, TEAD4, USF1, USF2, YY1). We find for the remaining 21 TFs that the motifs of the seven higher primates are more similar to each other compared to the motifs of dog, horse, and cow. With other words, in these cases the pairwise difference logos of dog, horse, and cow do not form a second cluster. The differences observed among species–specific motifs could partly result from missing binding sites in some species. **Supplementary Table S6** shows for each species and for each TF the proportion of sequences which are available for the computation of species–specific motifs.

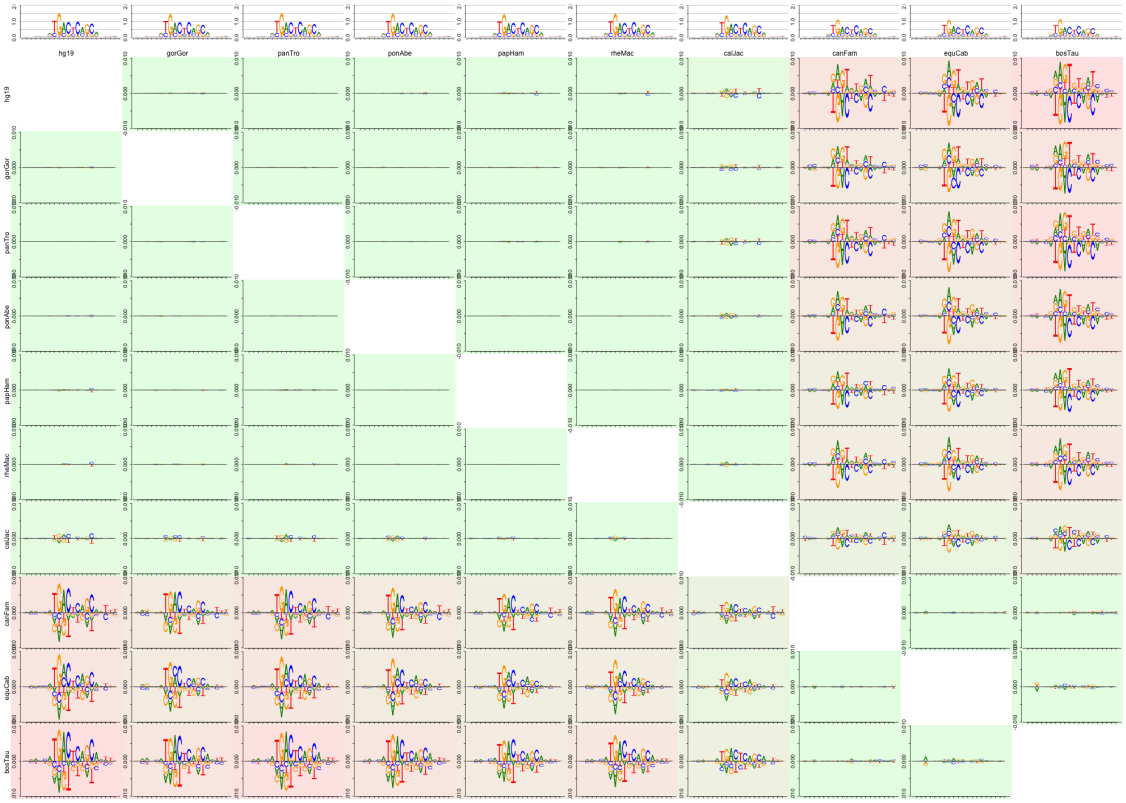


Figure 7.7: Comparison of species-specific motifs for the TF Bach1. We depict a table of difference logos with one row and one column for each species-specific motif to emphasize the differences between species-specific motifs. Each difference logos depicts the motif differences position-wise with a stack of bases, which height is calculated by the Jensen-Shannon divergence of the position-specific base distributions. The overall similarity between species-specific motifs is calculated by the sum of Jensen-Shannon divergences of all motif positions and depicted by the background color of the difference logos from green (similar) to red (dissimilar). The table of difference logos indicates, that the ten species-specific Bach1 motifs primarily segregate into two clusters, where one cluster comprises seven primates and the other cluster comprises three non-primates (cow, dog, and horse).

7.3.1.2 Species–specific motifs are typically highly similar

We study the statistical significance of differences between species–specific motifs as follows. We examine for each of the 35 TFs the similarity of species–specific motifs using a statistical test for each two species ($(10 * 9)/2 = 45$ species pairs). We calculate for each TF and for each two species the p-value for the null hypothesis that two species–specific motifs arise from the same distribution as described in **Supplementary Section 2.2** resulting in $35 * 45 = 1575$ pairwise comparisons. We count for each two species how often we reject the null hypothesis for a confidence level of $\alpha = 0.05$. These counts range from 0 to 35,

where 0 means that the species-specific motifs of two species show no significant differences for each transcription factor and 35 means that the species-specific motifs of two species show significant differences for each transcription factor. The binary nature of the results of statistical tests can lead to the issue that comparisons between three species are not transitive, i.e., if there are no significant differences between the species–specific motifs of species A and B and there are no significant differences between the species–specific motifs of species B and C it can happen that there are significant differences between the species–specific motifs of species A and C.

Supplementary Table S3 shows the results for each pair of species. We find that the seven primate-specific motifs are highly similar to each other and that the three species-specific motifs of cow, dog, and horse show greater differences compared to those of the seven primates. Using a significance level of 95%, we expect 5% of all 1575 pairwise differences to be significant by chance. We find for only 47 of 1575 pairwise comparisons that two species–specific motifs show significant differences. However, we find only 47 (3%) of the pairwise differences to be significant, stating that the observed differences are not greater than expected by chance.

Specifically, we find that these 47 cases apply only to 6 of the 35 TFs, namely Bach1, CEBPB, MafK, Max, SP1, and USF1 and typically only apply to comparisons between a primate species and one of the species dog, cow, and horse. We find no significant differences between the seven primates reflecting the close phylogenetic relationship and accordingly the high sequence similarity. Amongst the three species dog, cow, and horse we find for 1 of the 105 pairwise comparisons significant differences.

These results imply that the motifs estimated across species as presented in the previous section are typically not a mixture of species–specific motifs.

7.3.1.3 Intra–motif dependencies are highly similar for all species

We examine for each of the 35 TFs the distribution of species–specific MIs using mutual information profiles I_1^S and I_2^S as described in **Methods 4** for each species $S \in \{hg19, panTro, papHam, ponAbe, rheMac, calJac, equCab, canFam, gorGor, bosTau\}$. **Figure 7.8** shows two examples of species–specific mutual information profiles I_1^S and I_2^S for the two TFs CJUN and Nrf. All species–specific mutual information profiles are available in **Additional File 2**.

First, we study the species–specific mutual information profiles I_1^S. We find for each of the 35 TFs that the species–specific mutual information profiles I_1^S are highly similar for all species. We also find that the MIs in the mutual information profiles I_1 are sometimes stronger, sometimes weaker, and often averaged compared to the species–specific MIs I_1^S

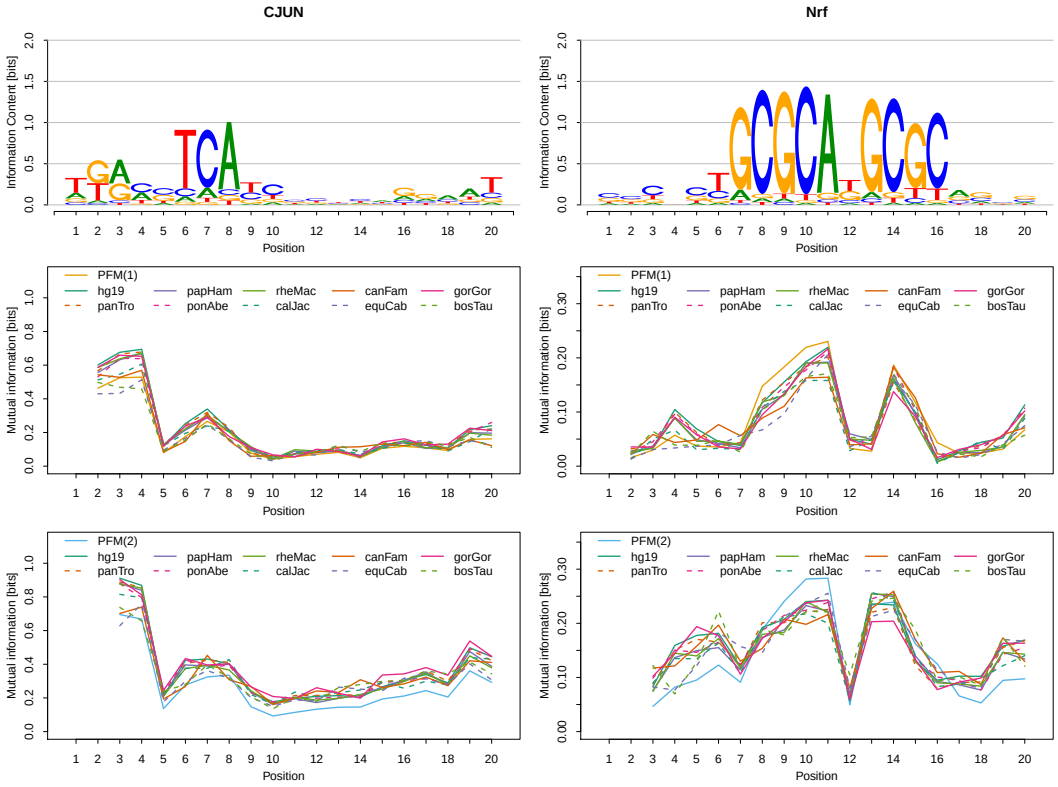


Figure 7.8: Sequence logos and intra-motif dependencies for the TFs (left) CJUN and (right) Nrf. We depict for both TFs (i) the sequence logo inferred by the PFM(2) from all species in the first row, (ii) the species-specific mutual information profiles inferred from the PFM(1) in the second row, and (iii) the species-specific mutual information profiles inferred from the PFM(2) in the third row. The species-specific mutual information profiles inferred from both models are highly similar to each other.

implying that the mutual information profiles I_1 inferred from all species are partly a result of interference of species–specific MIs. For example, in case of TF CJUN at motif positions $w \in \{2, 3, 4\}$, the MIs $I_1(w)$ are smaller than the MIs $I_1^S(w)$ for all species except horse and marmoset and in case of TF Nrf at motif positions $w \in \{8, 9, 10, 11\}$ the MIs $I_1(w)$ are higher than the MIs $I_1^S(w)$ for all species. Specifically, we find the largest difference between two I_1^S for FOSL1 with 0.35 bits. However, the mutual information profiles I_1 inferred from all species are typically highly similar to the species–specific mutual information profiles I_1^S.

Second, we examine species–specific mutual information profiles I_2^S. We find for each of the 35 TFs that the species–specific mutual information profiles I_2^S are highly similar for all species. We also find that the MIs in the mutual information profiles I_2 are sometimes stronger, sometimes weaker, and sometimes averaged compared to the species–specific MIs

I_2^S implying that the mutual information profiles I_2 inferred from all species are partly a result of interference of species–specific MIs. For example, in case of CJUN the mutual information profile $I_2(w)$ is typically smaller than the mutual information profile $I_2^S(w)$ for all species at all motif positions w and in case of Nrf at motif positions $w \in \{8-11\}$ the MIs $I_2(w)$ are higher than the MIs $I_2^S(w)$. Specifically, we find the largest difference between two I_2^S for FOSL1 with 0.48 bits. However, the mutual information profiles I_2 inferred from all species are typically highly similar to the species–specific mutual information profiles I_2^S as in case of I_1 and I_1^S.

7.3.2 Taking into account phylogeny improves classification performance in almost all cases.

It has been shown that taking into account base dependencies improves one-species approaches neglecting phylogenetic dependencies and it has been shown that taking into account phylogenetic dependencies can improve one-species approaches neglecting base dependencies. In the manuscript we have shown that taking into account base dependencies improves phylogenetic footprinting. Unfortunately, it can not be concluded from these observations that a model taking into account base dependencies and phylogenetic dependencies outperforms a model taking into account base dependencies but neglecting phylogenetic dependencies, because phylogenetic dependencies may potentially impair the model taking into account base dependencies.

Here, we systematically study the impact of both higher order base dependencies and phylogenetic dependencies to classification performance. Therefore, we study the performances of four different models, namely **i)** a model taking into account neither base dependencies nor phylogenetic dependencies (**human(0)**), **ii)** a model taking into account base dependencies of order 2 and neglecting phylogenetic dependencies (**human(2)**), **iii)** a model neglecting base dependencies and taking into account phylogenetic dependencies **PFM(0)**, and **iv)** a model taking into account both base dependencies and phylogenetic dependencies (**PFM(2)**) as described in **Methods 2**. The models PFM(0) and PFM(2) take into account phylogenetic dependencies and are inferred from the alignments described in **Methods 1**. The models human(0) and human(2) do not take not into account phylogenetic dependencies and are inferred from the human sequences of the alignments described in **Methods 1**. The models human(0) and human(2) are special cases of PFM(0) and PFM(2) incorporating only one species.

Based on these four models we perform all pair–wise comparisons, namely a) human(0) against human(2), b) human(0) against PFM(0), c) PFM(0) against PFM(2), d) human(2) against PFM(2), e) human(0) against PFM(2), and f) human(2) against PFM(0).

For case a), it has been shown that human(2) typically outperforms human(0), i.e., that modeling base dependencies improves classification performance. For case b), it has also been shown that PFM(0) typically outperforms human(0), i.e., modeling phylogenetic dependencies improves classification performance. For case c), we have shown that PFM(2) outperforms PFM(0), i.e., that taking into account higher order base dependencies improves phylogenetic footprinting. For case e) we assume that PFM(2) outperforms human(0) considering the cases a) and b). The cases d) and f) are unknown so far.

We measure the classification performance of all four models as described in **Methods 3** on datasets of 35 TFs. **Figure 7.9** shows the corresponding values for the four models human(0), PFM(0),human(2), and PFM(2) for each of the 35 TFs. See **Supplementary**

Table S4 and **Supplementary Table S5** for statistics of the results shown in **Supplementary Figure 7.9**.

It is not surprising that the model taking into account both base dependencies and phylogenetic dependencies outperforms the model ignoring base dependencies and phylogenetic dependencies (case e). We find that modeling base dependencies typically improves classification performance (cases a and c). Interestingly, we find that modeling base dependencies clearly outperforms modeling phylogenetic dependencies (case f). In fact, solely taking into account phylogenetic dependencies shows only a partial improvement (case b), but taking into account both phylogenetic dependencies and base dependencies shows a clear performance improvement compared to solely taking into account base dependencies (case d). These results suggest that phylogenetic footprinting approaches benefit from taking into account base dependencies and that approaches on single species already taking into account base dependencies benefit from taking into account phylogenetic dependencies.

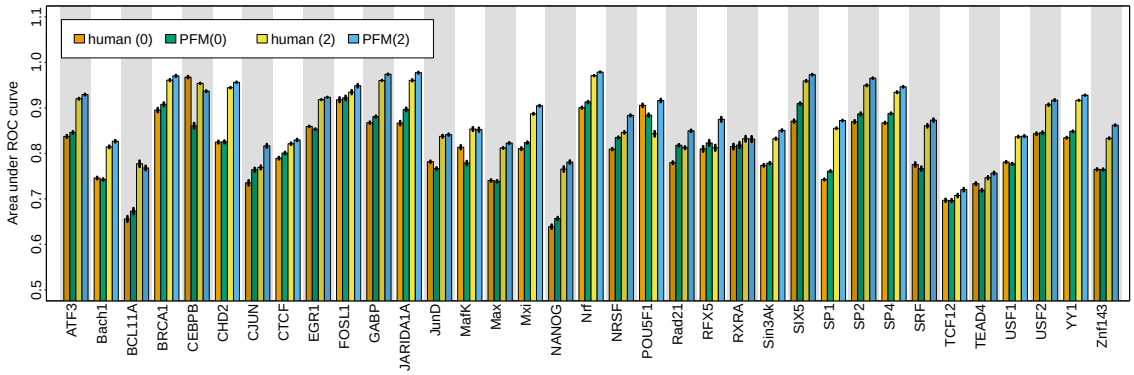


Figure 7.9: Classification performance of the models human(0) and human(2) on human sequences and PFM(0) and PFM(2) on alignments of ten species for each of the 35 **TFs.** We show the mean and standard error of the ROC AUC. See **Table S4** and **Table S5** for summary statistics.

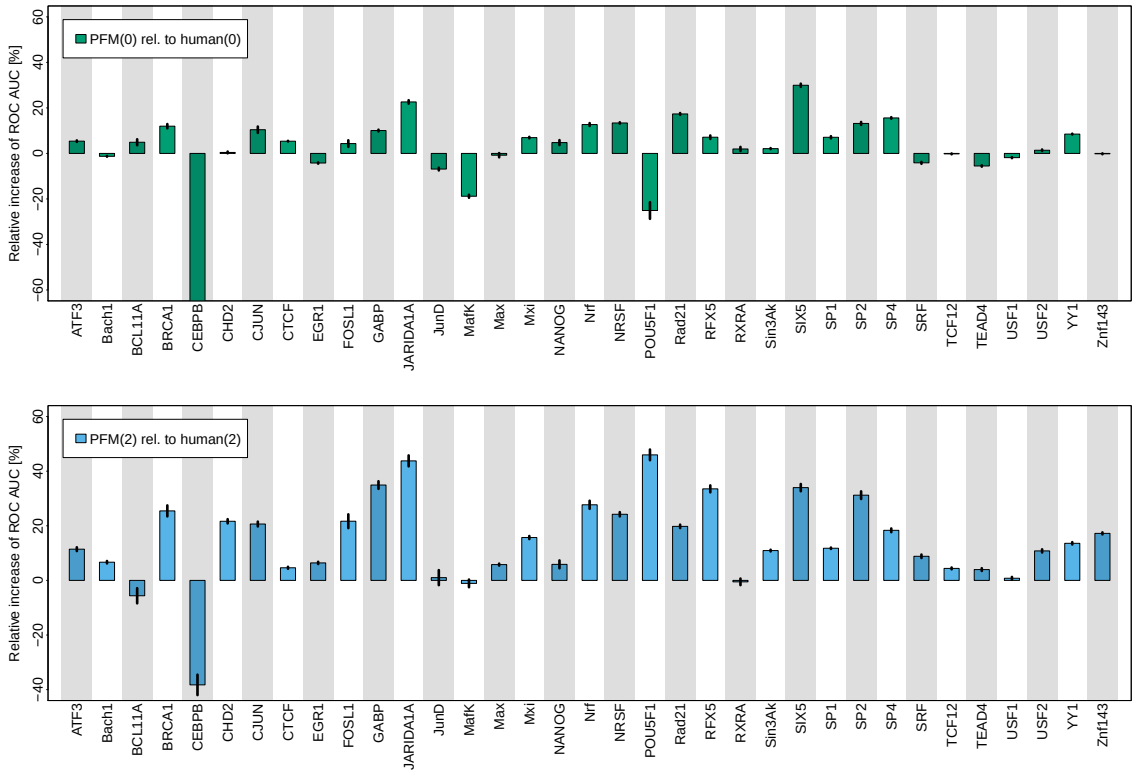


Figure 7.10: Classification performance of the two models PFM(0) and PFM(2) incorporating phylogenetic dependencies for each of the 35 **TFs.** We show the mean and standard error of the relative increase of ROC AUC of **(top)** PFM(0) relative to the classification performance of human(0) and **(bottom)** PFM(2) relative to the classification performance of human(2). Typically both models show a higher classification performance.

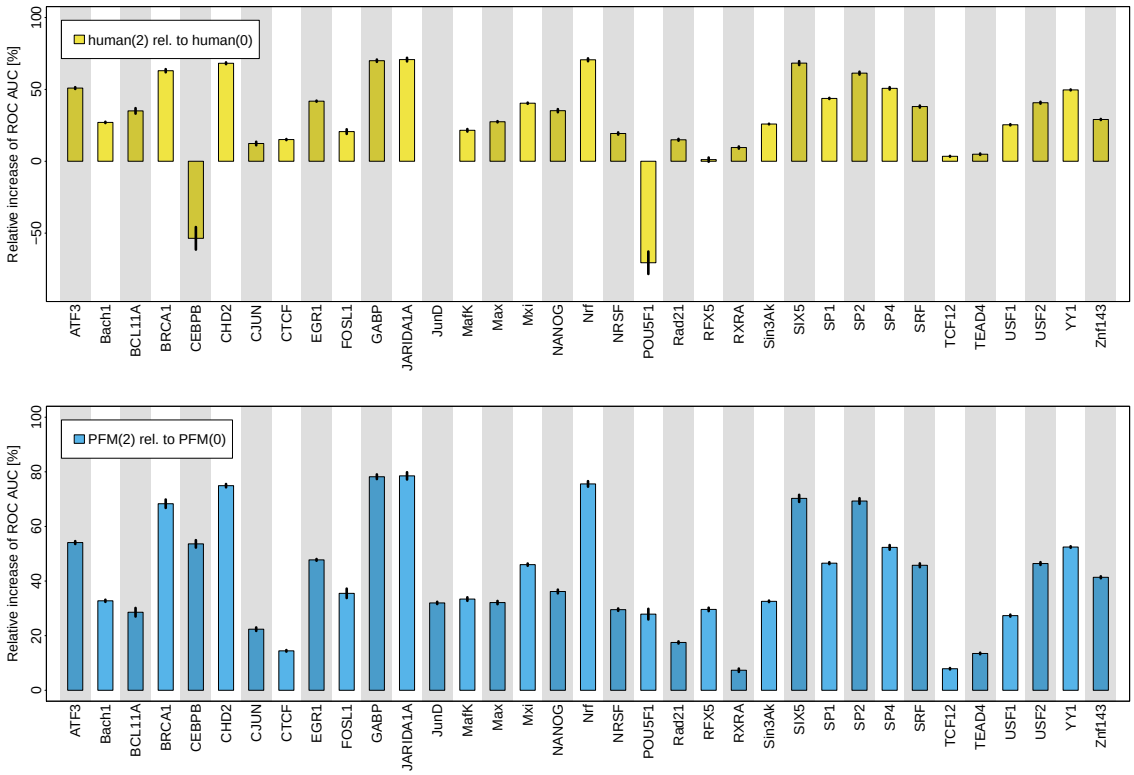


Figure 7.11: Classification performance of the two models human(2) and PFM(2) incorporating base dependencies of order two for each of the 35 TFs. We show the mean and standard error of the relative increase of ROC AUC of **(top)** human(2) relative to the classification performance of human(0) and **(bottom)** PFM(2) relative to the classification performance of PFM(0). Typically both models show a higher classification performance.

7.4 DiffLogo: A comparative visualisation of sequence motifs

The supplementary material of "DiffLogo: a comparative visualization of sequence motifs" consists of one additional file that contains Supplementary Methods, Results, Figures, and Examples. This file comprises four sections. **Section 1**, *Additional examples*, contains **Figures S1** and **S2**. In **Section 2**, CTCF with and without clustering, we show in detail the impact of clustering and optimal leaf ordering for a DiffLogo grid of nine CTCF motifs. In **Section 3**, *Alternative combinations of stack heights and symbol weights*, we describe the mathematical background of four implementations of H_ℓ and two implementations of $r_{\ell,a}$ and show an exemplary comparison of the eight combinations. In **Section 4**, *Tool comparison*, we compare DiffLogo with the five tools seqLogo, iceLogo, MotifStack, STAMP, and Two Sample Logo. Here, I provide a copy of **Section 3**.

7.4.1 Alternative combinations of stack heights and symbol weights

We consider two motifs represented by two PWMs p and q. The height of symbol a in the symbol stack at position ℓ of the difference logo is denoted $H_{\ell,a}$ and given by

$$H_{\ell,a} = r_{\ell,a} \cdot H_\ell,$$

where H_ℓ represents the height of the symbol stack at position ℓ and the weight $r_{\ell,a}$ represents the proportion of symbol $a \in \mathcal{A}$ in the symbol stack at position ℓ, where $\mathcal{A}$ is the alphabet. We calculate $H_{\ell,a}$ for different measures H_ℓ and $r_{\ell,a}$ to emphasize different facets of distribution differences. We propose various alternatives to calculate the measures H_ℓ and $r_{\ell,a}$ as follows (illustrated in supplementary Table S1.

In the following sections, the information content of a PWM p at position ℓ is denoted H_ℓ^p and given by

$$H_\ell^p = \log_2(|\mathcal{A}|) - \sum_{a \in \mathcal{A}} p_{\ell,a} \cdot \log_2(p_{\ell,a}),$$

where $p_{\ell,a}$ is the probability of symbol a at position ℓ in PWM p. H_ℓ^q is defined analogously.

7.4.1.1 Different calculations of stack heights H_ℓ

Jensen–Shannon divergence

The Jensen–Shannon divergence is a measure for the difference of two probability distributions based on information theory. The Jensen–Shannon divergence at position ℓ is denoted by $H_\ell^{(\mathrm{i})}$ and given by

$$\begin{aligned} H_\ell^{(\mathrm{i})} =& \frac{1}{2} \sum_{a \in \mathcal{A}} p_{\ell,a} \Big(\log_2(p_{\ell,a}) - \log_2(m_{\ell,a}) \Big) \\ &+ \frac{1}{2} \sum_{a \in \mathcal{A}} q_{\ell,a} \Big(\log_2(q_{\ell,a}) - \log_2(m_{\ell,a}) \Big), \end{aligned}$$

where $m_{\ell,a} = \frac{1}{2}(p_{\ell,a} + q_{\ell,a})$. $H_\ell^{(\mathrm{i})}$ is symmetric and limited to $[0, 1]$. This measure especially emphasizes large distribution differences.

Change of information content (stack)

The change of information content (stack) is a measure for the absolute change of information content between two probability distributions. The change of information content (stack) at position ℓ is denoted by $H_\ell^{(\mathrm{ii})}$ and given by

$$H_\ell^{(\mathrm{ii})} = \sum_{a \in \mathcal{A}} |p_{\ell,a} H_\ell^p - q_{\ell,a} H_\ell^q|.$$

$H_\ell^{(\mathrm{ii})}$ is symmetric and limited to $[0, 2 * \log_2(|\mathcal{A}|)]$. This measure especially emphasizes large changes of information content.

Relative change of information content

The relative change of information content is a measure for the absolute change of information content relative to the average information content of the two probability distributions. The relative change of information content at position ℓ is denoted by $H_\ell^{(\mathrm{iii})}$ and given by

$$H_\ell^{(\mathrm{iii})} = \begin{cases} \dfrac{\sum_{a \in \mathcal{A}} |p_{\ell,a} H_\ell^p - q_{\ell,a} H_\ell^q|}{\frac{1}{2}(H_\ell^p + H_\ell^q)} & \text{if } p_\ell \neq q_\ell \\ 0 & \text{otherwise.} \end{cases}$$

$H_\ell^{(\mathrm{iii})}$ is symmetric and limited to $[0, 2 * \log_2(|\mathcal{A}|)]$. This measure especially emphasizes large changes of information content relative to the information content of the given distributions.

Change of probabilities (stack)

The change of probabilities (stack) is a measure for the absolute change of probabilities between two probability distributions. The change of probabilities (stack) at position ℓ is denoted by $H_\ell^{(\mathrm{iv})}$ and given by

$$H_\ell^{(\mathrm{iv})} = \sum_{a \in \mathcal{A}} |p_{\ell,a} - q_{\ell,a}|$$

$H_\ell^{(\mathrm{iv})}$ is symmetric and limited to $[0, 2]$. This measure especially emphasizes large changes of probabilities.

7.4.1.2 Different calculations of symbol weights $r_{\ell,a}$

Change of probability (symbol)

The change of probability (symbol) is a measure for the change of symbol-specific probability relative to the sum of absolute symbol-specific probability differences of the given probability distributions. The change of probability (symbol) of symbol a at position ℓ is denoted by $r_{\ell,a}^{(\mathrm{i})}$ and given by

$$r_{\ell,a}^{(\mathrm{i})} = \begin{cases} \frac{p_{\ell,a} - q_{\ell,a}}{\sum_{a' \in \mathcal{A}} |p_{\ell,a'} - q_{\ell,a'}|} & \text{if } p_\ell \neq q_\ell \\ 0 & \text{otherwise.} \end{cases}$$

$r_{\ell,a}^{(\mathrm{i})}$ is antisymmetric and limited to $[-\frac{1}{2}, \frac{1}{2}]$. This measure especially emphasizes a large change of symbol–probability. For each position of the difference logo, the height of the symbol stack with negative measures $r_{\ell,a}^{(\mathrm{i})}$ is equal to the height of the symbol stack with positive measures $r_{\ell,a}^{(\mathrm{i})}$, because each gain of symbol–probability implies a loss of probability for the remaining symbols and vice versa.

Change of information content (symbol)

The change of information content (symbol) is a measure for the symbol-specific change of information content relative to the sum of absolute symbol-specific differences of information content of the given probability distributions. The change of information content (symbol) of symbol a at position ℓ is denoted by $r_{\ell,a}^{(\mathrm{ii})}$ and given by

$$r_{\ell,a}^{(\mathrm{ii})} = \begin{cases} \frac{p_{\ell,a} H_\ell^p - q_{\ell,a} H_\ell^q}{\sum_{a \in \mathcal{A}} |p_{\ell,a} H_\ell^p - q_{\ell,a} H_\ell^q|} & \text{if } p_\ell \neq q_\ell \\ 0 & \text{otherwise.} \end{cases}$$

$r_{\ell,a}^{(\mathrm{ii})}$ is antisymmetric and limited to $[-1, 1]$. This measure especially emphasizes a large change of symbol-specific information content.

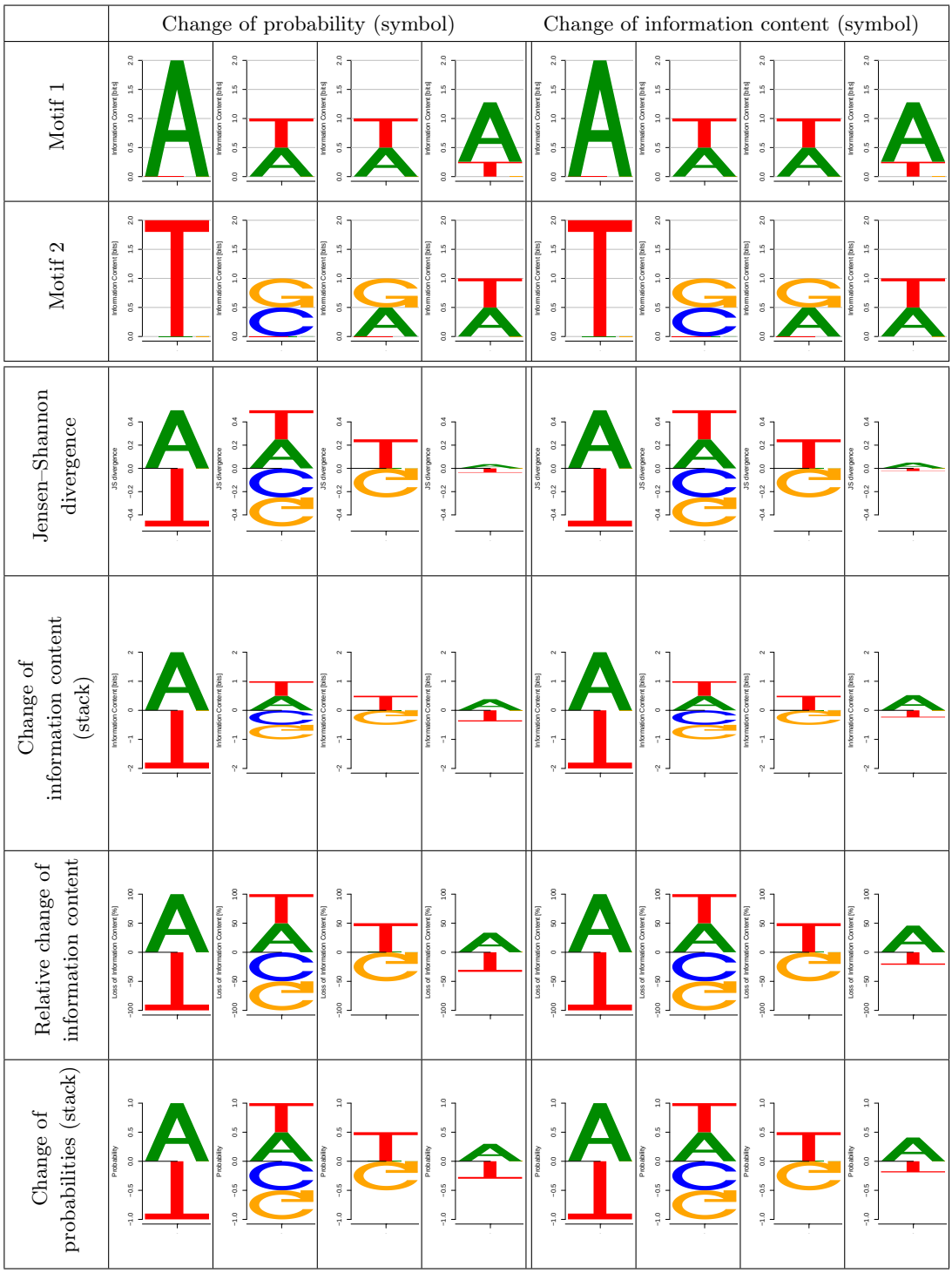


Table 7.1: Exemplary comparison of different stack heights and symbol weights using four artificial DNA motifs of length one.

Eidesstattliche Erklärung / Declaration under Oath

Ich erkläre an Eides statt, dass ich die Arbeit sebstständig und ohne fremde Hilfe verfasst, keine anderen als die von mir angegegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

I declare under penalty of perjury that this thesis is my own work entirely and has been written without any help from other people. I used only the sources mentioned and included all the citations correctly both in word or content.

Burgliebenau, 15.05.2017 Martin Nettling

Martin Nettling

Phone: +49 173 360 12 83 | mgledi@gmail.com

Personal information

Address
Die Mühlbreite 21
06258 Burgliebenau
Date of birth, birth name
10.06.1982, Gleditzsch
Nationality
German
Family status
married, two kids

Education

07/2008 **Diploma Bioinformatics**
Martin-Luther-University, Halle/Wittenberg

06/2001 **Abitur**
European school in Waldenburg

Professional experience

01/2016 – today
Senior Software Engineer Research and Development
Datameer GmbH, Halle
» Development of distributed algorithms
» Development of Elasticsearch plugins

08/2012 – 10/2015
Teamlead Research and Development
Unister GmbH, Leipzig
» Analysis and interpretation of large data sets
» Development of distributed algorithms
» Leading a team of up to six team members

10/2010 – 07/2012
Junior Developer Machine-Learning
Unister GmbH, Leipzig
» Implementation of algorithms for text analysis
» Performance optimization of existing implementations

11/2003 – 06/2007
Developer of the CMS of the MLU Halle/Wittenberg
» Conception and backend development

Martin Nettling

Phone: +49 173 360 12 83 | mgledi@gmail.com

Publications

Articles

M Nettling, H Treutler, J Cerquides, I Grosse. 2016. Unrealistic phylogenetic trees may improve phylogenetic footprinting. *Bioinformatics, accepted*

M Nettling, H Treutler, J Cerquides, I Grosse. 2016. Combining phylogenetic footprinting with motif models incorporating intra-motif dependencies.
BMC bioinformatics, accepted

M Nettling, H Treutler, J Cerquides, I Grosse. 2016. Detecting and correcting the binding-affinity bias in ChIP-Seq data using inter-species information. *BMC genomics* 17:1.

M Nettling*, H Treutler*, J Grau, J Keilwagen, S Posch, I Grosse. 2015. DiffLogo: a comparative visualization of sequence motifs. *BMC Bioinformatics*, 16:1

M Nettling, N Thieme, A Both, I Grosse. 2014. DRUMS: Disk Repository with Update Management and Select option for high throughput sequencing data. *BMC bioinformatics*, 15:1.

P Alexiou, T Vergoulis, **M Gleditzsch**, G Prekas, T Dalamagas, M Megraw, I Grosse, T Sellis, AG Hatzigeorgiou. 2009. miRGen 2.0: a database of microRNA genomic information and regulation. *Nucl. Acids Res.* 38 (suppl 1): D137-D141

Conference papers

L Avdiyenko, **M Nettling**, C Lemke, M Wauer, ACN Ngomo, A Both. (2015) Motive-based search: Computing regions from large knowledge bases using geospatial coordinates. *International Joint Conference on Knowledge Engineering and Knowledge Management (IC3K)*, (Vol. 1, pp. 469-474). SCITEPRESS.

M Wauer, A Both, S Schwinger, **M Nettling**, O Erling. (2015) Integrating custom index extensions into virtuoso RDF store for e-commerce applications. *Proceedings of the 11th International Conference on Semantic Systems* (pp. 65-72). ACM.

F Rosner, A Hinneburg, M Röder, **M Nettling**, A Both. (2013) Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397*.

F Rosner, A Hinneburg, **M Gleditzsch**, M Priebe, A Both. (2012) Fast sampling word correlations of high dimensional text data. *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (pp. 866-866). ACM.

Martin Nettling

Phone: +49 173 360 12 83 | mgledi@gmail.com

Open source software

- **DRUMS:** Disk repository with update management and select option
  https://github.com/mgledi/DRUMS
- **BioDRUMS:** Example integration of DRUMS for biological data
  https://github.com/mgledi/BioDRUMS
- **DiffLogo:** Comparative visualization of sequence motifs in R
  https://github.com/mgledi/DiffLogo
- **DiffLogoUI:** A webserver for DiffLogo
  https://github.com/mgledi/DiffLogoUI

Burgliebenau, 01.12.2016 Martin Nettling