Computational Approaches for Analyzing the Role of Protein-DNA Interactions in Gene Regulation

Dissertation
der Fakultät für Informations- und Kognitionswissenschaften
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl.-Ing. Annette Höglund
aus Mariehamn, Finnland

Tübingen
2006

Tag der mündlichen Qualifikation:	19.06.2006
Dekan:	Prof. Dr. Michael Diehl
1. Berichtstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichtstatter:	Prof. Dr. Hans-Peter Lenhof (Universität des Saarlandes)
3. Berichtstatter:	Dr. Hagit Shatkay (Queen’s University, Kingston)

Acknowledgements

Standing here with a nice feeling inside, there are several people I want to acknowledge.

First and foremost, I want to thank Prof. Oliver Kohlbacher and Prof. Hans-Peter Lenhof for giving me the opportunity to choose an interesting research topic and for providing excellent conditions to explore it both in Tübingen and in Saarbrücken.

I am forever indebted to Oliver Kohlbacher for an enthusiastic supervision, stimulating discussions, and for taking me on as a PhD student in the first place. I admire him for a sharp mind, a fine sense to combine critique with commitment towards others, and for being a skilled team leader.

Looking further back in time I want to start with thanking Arne Elofsson at the Stockholm Bioinformatics Center, for showing me what computer science can do for life sciences. Realizing that is my mere motivation for staring at something so incredibly square-shaped for so incredibly many hours.

I was fortunate to meet Wyeth Wasserman, who warmly welcomed me into his group at the Karolinska Institute and opened the door to future research. Albin Sandelin, Boris Lenhard, Nina Ståhlberg, and Elena Herzog contributed with inspiring discussions, great work, and a wonderful group atmosphere.

Thanks to Gene Myers, Vineet Bafna, Samuel Levy, and Sridhar Hannenhalli for inviting me to do an internship at Celera Genomics. Knut Reinert and Daniel Huson equipped me with a weapon and protection in order to survive regular afternoon battles. Karen Eilbeck and Giuseppe Lancia provided housing, dancing company, and were never late for laughs. Nathan Edwards and Sorin Istrail for infectious mirth. Hagit Shatkay was happy to inherit my armour and is the most encouraging woman I know. Furthermore, I want to thank Hagit for her efforts in reviewing the work described in this thesis.

My warm thanks to Hans-Peter Lenhof and the rest of the very friendly Saarbrücken kickers, for their hospitality and support during the first few years in Germany.

Much respect to my current colleagues in Tübingen: Andreas Kerzmann, Pierre Dönnes, Marc Sturm, Torsten Blum, Muriel Quenzer, Jana Schmidt, Nora Toussaint, and Nico Pfeifer for discussions, thinking, company, care, and assistance with all kinds of technical problems

at all times. Thanks to my former and present students for their hard work, inspiration, and patience. I specifically want to emphasis the work by Marc Sturm (development and implementation of CAP), Christine Herold (analysis of human melanoma-related genes), Sebastian Schultheiss (analysis of stem cell regulation in plants), and Nina Fischer (for her stubbornness and shared interest in modeling 3D structures). Furthermore, I want to thank the members of the Zentrum für Bioinformatik Tübingen (ZBIT) for contributing to enjoyable moments at Sand and in Oberjoch.

I want to thank Hagit Shatkay, Stefan Rensing, Jan Lohmann, Wolfgang Busch, Hans-Werner Adolph, Jens Lagergren, Michael Hallett, Michelle Scott, and Martin Latterich for pleasant cooperations.

I am grateful beyond what words can describe to my parents Ann-Louise and Henry Höglund for endless encouragement and concern, boosting my self-esteem, and for being just what they are - my parents. Thanks to my brothers, grandparents, and Pierre's family for reminding me about life in general, care, and attention whenever it is needed.

A warm bunch of hugs to all my friends for support, entertaniment, and for trying to keep me sane. In particular, I want thank to my best friend Johanna Gustafsson for always being a great source of strength and for kindling an eternal spirit of optimism.

Emmy

our wonderful daughter, for showing me a new dimension to life. Your sparkling smile at the crack of dawn reminds me of how beautiful life is - every day of it!

Pierre

my everything. You hear my thoughts, understand my dreams, and fill my life with laughter and love!

Abstract

Gene regulation plays a pivotal role at all stages of organism development, in cell differentiation, and for maintaining homeostasis. Controlled spatial and temporal gene expression is achieved by means of complex and robust regulatory networks. A key event in maintaining such networks is the sequence specific protein-DNA recognition, which enables transcription factors to identify their respective binding sites.

Computational and structural biologists face intriguing challenges at three different levels when investigating gene regulation. First, the involvement of gene regulation in disease can be addressed by studying global effects of gene regulatory networks, which are visible at the level of systems. Furthermore, detecting the often short and variable transcription factor binding sites (TFBSs) in genomic DNA is not a trivial task, since the prediction of TFBSs and delineation of functional regulatory modules are conducted at the level of sequences. Finally, there is a challenge in understanding the factors governing transcription factor-DNA recognition, as the information needs to be collected at the molecular level. Structure-based methods provide detailed information about protein-DNA interactions at atomic resolution.

In this work, a versatile approach for computational analysis of the different levels of gene regulation, gradually zooming in from the global level of systems to the molecular level, is presented. Linking information related to gene regulation from the different levels can help in clarifying phenomena that are hard to explain using only one source of information. First, the influence of gene regulation is analyzed at the level of systems. A set of cancer-related target genes are identified using a novel integrative analysis pipeline. Microarray data, immunological data, and curated biological knowledge are brought together enabling extensive analysis of the underlying mechanisms controlling gene expression in cancer tissue. The transcription factor AP2 is suggested to play a key regulatory role in controlling a set of over-expressed melanoma-related genes. The computational results presented are supported by previously reported experimental evidence.

Zooming in to the level of sequences transcription factors orchestrating the expression of functionally related genes are identified in yeast and plant, which are two important model organisms for studying gene regulation. The pattern-finding algorithm Gibbs sampling is employed for discovering putative functional TFBSs in functionally related genes. The response element *ACGCGT* is found to be over-represented in DNA-repair genes in yeast, which supports the idea that the transcription factor MBP1 is involved in blocking repli-

cation of damaged DNA. The vital regulation of stem cells is explored in plant, providing preliminary computational evidence for TFBSs critical to stem cell differentiation.

The final transition is the step from analyzing gene regulation at the levels of systems and sequences to studying protein-DNA interactions at atomic detail. Structural data provides an additional source for gaining insight into the thermodynamic properties of sequence specific binding, which eventually directs gene regulation. A computational protocol for analyzing the effects that small base modifications have on the overall binding free energy is described. The computationally obtained results for mutating the thymine to uracil in transcription factor-DNA complexes agree well with previously reported experimental results, illustrating the applicability of the protocol. This is a first step towards using molecular modeling for constructing structure-based models of TFBSs.

Each individual level of this step-wise analysis provides crucial information needed to gain insight into the different aspects underlying complex regulatory control mechanisms. Analysis at the level of systems and networks is crucial for understanding global effects of gene regulation, the implications of gene regulation in disease, and for identifying sets of target genes. Sequence-based methods are used for discovering functional binding sites in gene regulatory regions for such sets of related genes, responsible for directing gene expression. Finally, structural analysis can explain ambiguities observed in sequence-based models, however, can only be applied to a limited number of protein-DNA complexes due to high computational requirements. An improved understanding of all aspects of gene regulation is inevitable for identifying key factors influencing organism development and disease.

Kurzzusammenfassung

Genregulation spielt eine entscheidende Rolle in allen Entwicklungsstadien eines Organismus, bei der Zelldifferenzierung und dem Erhalt der Homöostase. Die kontrollierte räumliche und zeitliche Expression bestimmter Gene wird dabei durch ein komplexes, aber robustes, Netzwerk kontrolliert. Ein Schlüsselprozess der Regulation ist dabei die sequenzspezifische Protein-DNA-Erkennung, die es Transkriptionsfaktoren erlaubt ihre jeweiligen Bindungsstellen zu erkennen.

Die Untersuchung der Genregulation wirft interessante Fragen auf drei verschiedenen Ebenen auf. Auf der obersten Ebene, der Ebene der Systeme, beschäftigt man sich dabei mit den Auswirkungen der Genregulation auf Netzwerke als Ganzes. Diese Ebene hat wichtige Implikationen für die Erforschung von Krankheiten. Die zweite Ebene, die Sequenzebene, betrachtet die Wechselwirkungen von Transkriptionsfaktoren mit ihren genomischen Bindestellen und erlaubt Aussagen über regulatorische Module und deren Anordnung im Genom. Die dritte, molekulare Ebene schließlich versucht die Protein-DNA-Wechselwirkungen ausgehend von der dreidimensionalen Struktur von Proteinen und DNA zu erklären.

In dieser Arbeit werden eine Reihe von Ansätzen zur rechnergestützten Analyse der Genregulation auf all diesen Ebenen vorgestellt, von der Systemebene bis hinab zur molekulare Ebene. Zunächst wird dabei der Einfluss der Genregulation auf der Systemebene betrachtet. Mit einer neuen integrativen Analyse-Pipeline werden dazu an der Entstehung von Krebs beteiligte Gene identifiziert. Dazu wird eine ganze Reihe heterogener Datensätze integriert und im gemeinsamen Kontext analysiert, insbesondere in Bezug auf die Genexpression in Krebsgeweben. Es stellt sich heraus, dass der Transkriptionsfaktor AP2 eine Schlüsselrolle in der Steuerung überexprimierter Gene in Melanomen spielt. Diese theoretisch erhaltenen Ergebnisse unterstützen früher erzielte experimentelle Ergebnisse.

Geht man nun einen Schritt weiter hinab, zur Ebene der Sequenzen, so kann man hier an anderen Modellsystemen, Hefe und der Ackerschmalwand, das Zusammenspiel verschiedener Transkriptionsfaktoren in der Regulation funktionell verwandter Gene studieren. Mit Gibbs-Sampling wurden dazu potentielle Bindestellen von Transkriptionsfaktoren identifiziert. Dabei stellt sich insbesondere das response element ACGCGT als überrepräsentiert in regulatorische Regionen von DNA-Reparaturgenen der Hefe heraus. Dies unterstützt die Hypothese, dass der Transkriptionsfaktor MBP1 beim Blockieren der Replikation beschädigter DNA beteiligt ist. In *Arabidopsis thaliana*, der Ackerschmalwand, wurde mit ähnlichen Methoden die Re-

gulation der pflanzlichen Stammzellen untersucht. Vorläufige Ergebnisse deuten hier auf die kritischen Rollen bestimmter Transkriptionsfaktoren hin und leisten einen Beitrag zur Aufklärung der zugrunde liegenden regulatorischen Netzwerke.

Geht man schließlich eine weitere Ebene hinab, so kann man die Interaktion der Transkriptionsfaktoren mit der DNA auf molekularer Ebene untersuchen. Ausgehend von strukturellen Daten von DNA-Transkriptionsfaktor-Komplexen lassen sich die thermodynamischen Größen bestimmen, die für die Regulation ausschlaggebend sind. Es wird ein Simulationsprotokoll vorgestellt, dass es erlaubt, den Einfluss von Punktmutationen in der DNA auf die freie Bindungsenthalpie zu berechnen. Die derart bestimmten Änderungen der freien Enthalpie für Mutationen von Thymin zu Uracil in Zinkfinger-DNA-Komplexen stimmen sehr gut mit experimentell bestimmten Werten überein. Diese Art von Studien ist ein erster Schritt zur Vorhersage der Motive eines Transkriptionsfaktors ausgehend von seiner Struktur.

Eine solche Sicht auf die verschiedenen Ebenen des Phänomens Genregulation erlaubt ein besseres Verständnis des gesamten Vorgangs. Jede Ebene liefert wesentliche Informationen zu einem bestimmten Aspekt der Genregulation: die systemische Ebene erlaubt das Verständnis der Regulation im Kontext des gesamten regulatorischen Netzwerks und erlaubt es, die Effekte der Genregulation auf komplexe Krankheitsverläufe zu untersuchen. Sequenzbasierte Methoden erlauben das Verständnis der lokalen Feinregulation funktionell verwandter Gene. Die molekulare Ebene schließlich erlaubt es, die Mehrdeutigkeiten sequenzbasierter Modelle zu verstehen und vorherzusagen. Der hohe Rechenaufwand dieser Methoden beschränkt diese Art von Modell aber derzeit noch auf kleine Studien und ausgewählte Beispielfälle.

Contents

1 Introduction

The coordinated interaction of biomolecules is essential for achieving specificity and efficiency of virtually all cellular processes. Gene regulation plays a central role in the cellular machinery as it determines cell differentiation and thereby organism development. Primarily controlled at the level of transcription, gene regulation is conducted by DNA-binding proteins known as transcription factors (TFs). Typically, a set of TFs act in concert to ensure correct spatial and temporal expression of a set of target genes [83]. The combinatorial aspect increases the degree of complexity and is a key feature for precisely controllable regulatory cascades and networks [163].

It is of outmost importance that the correct transcriptional control cascade is triggered, especially during the early stages of organism development and in response to external factors [156, 236]. Alterations in genes, proteins, or interactions between them can lead to dysfunctioning control and transport systems, which eventually can result in disease [6, 242]. Genetic aberrations occurring in genes coding for TFs can have wide-spread effects, due to their critical functions they exercise at all stages of organism development. Consequently, research efforts are typically motivated by specific biological and disease-related applications, aiming to understand the underlying causes and their implications [144].

Gene regulation can be studied at different levels, which all provide invaluable sources of information and guidelines for further research. The different aspects of gene regulation, ranging from the level of biological systems to atomic detail, are introduced in the following paragraphs. A graphical illustration of the different levels, at which gene regulation can be studied, is shown in Fig. 1.1.

Gene regulatory networks control gene expression observed at the level of systems. Hence, global techniques are necessary for assessing the implications of dysfunctioning proteins and effects of therapeutic agents. Systematic perturbation of genetic networks is performed for functional characterization [193, 112] and provides a useful tool in the drug target identification process [42]. Global gene expression profiles are often used for diagnostic purposes by

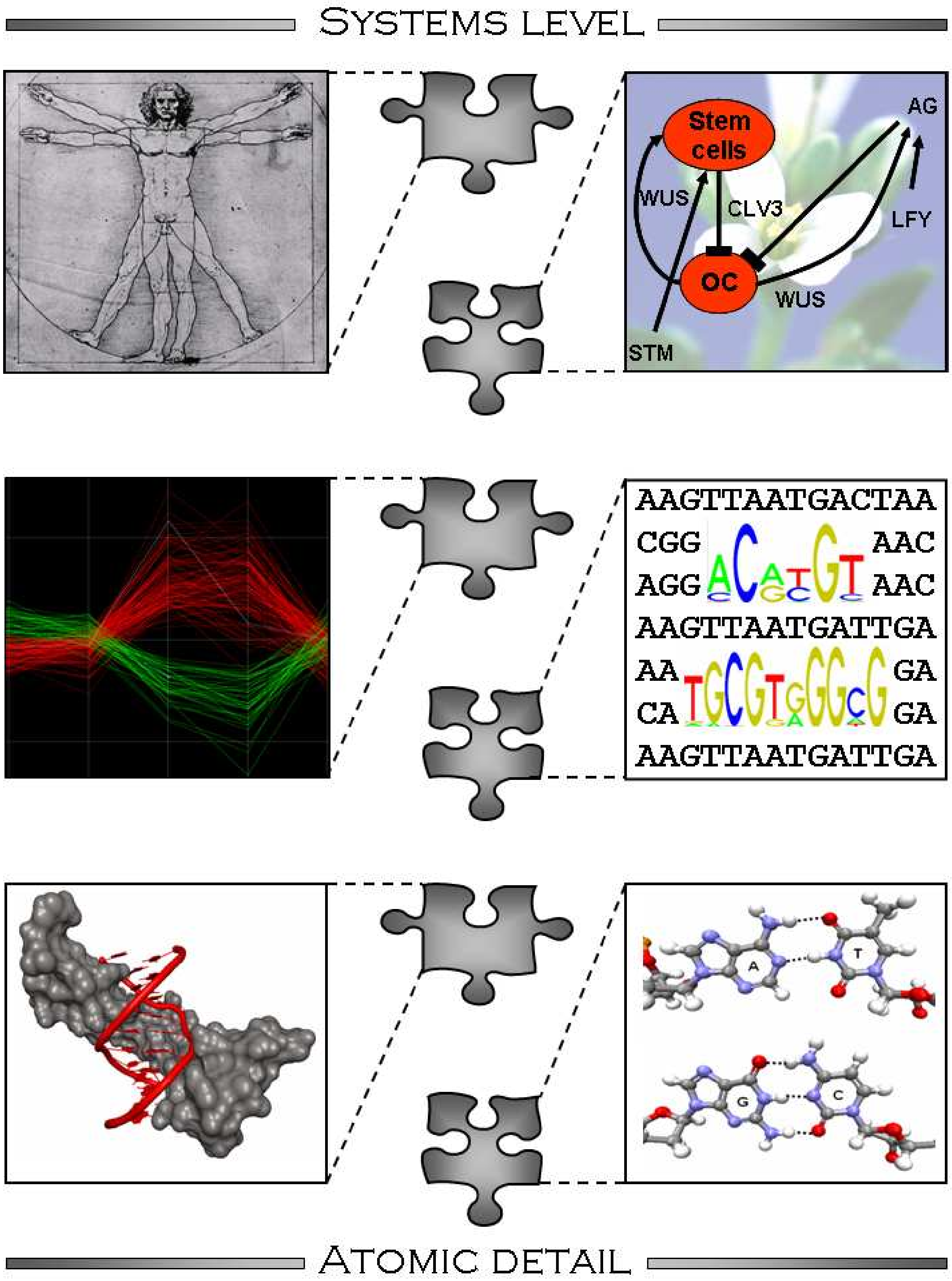


Figure 1.1: A graphical illustration demonstrating the different levels, at which gene regulation can be analyzed. Observing gene regulatory networks at the level of biological systems gives an overview of (top). Gene expression profiles are used for characterizing gene function and for identifying individual TFBSs at the level of gene regulatory sequences (middle), which is useful for reconstructing higher level networks. Zooming in further, gene regulation can be studied at atomic detail (bottom). Structural data of protein-DNA complexes provides insight into specific interactions governing protein-DNA association and has proven useful for explaining observations at the sequence level.

studying the behavior of marker genes e.g. for determining the disease progression or cancer type [13].

Understanding functional relationships between genes and constructing network models of their transcriptional regulation is a challenging task. A common way to address the individual components of regulatory networks is to analyze the involved genes at the sequence level. Sequence-based methods aim to delineate the functional transcription factor binding sites (TFBSs) within the relevant regulatory sequences [68] and allow for detailed studies of effects due to mutations. Relationships between key players involved in gene regulation can be elucidated by collecting gene expression data from several complementing experiments [112].

The sequence-based models of the binding sites for specific TFs are typically inadequate. Representing the short (and often variable) TFBSs achieving the high level of selectivity and sensitivity that TFs have, is a true challenge. In order to improve the models of the binding sites, the interactions underlying sequence specific recognition between proteins and DNA have been subject to structure-based studies at atomic detail [123, 166, 264]. Sequence specificity, which is the basis for transcriptional control is achieved through structural and chemical complementarity between a TF and its binding site.

The theoretical background, both biological and computational, needed for understanding the work in this thesis is presented in Chapt. 2. The concepts of transcriptional control of gene expression, the role and biochemical properties DNA-binding proteins, and the complexity of regulatory networks are introduced. Furthermore, the basic theoretical concepts of computational chemistry and molecular modeling are described. Related work, including both experimental and the most commonly employed computational approaches, is outlined in Chapt. 3.

Chapters 4 to 6 contain several studies, where protein-DNA interactions are the focus. The versatile approach presented here addresses three important aspects of gene regulation, which are illustrated in Fig. 1.1. Each of these levels are addressed individually - moving from systems biology (top) to the atomic level analysis (bottom) - and the results are considered from a more global point of view. First, the influence of transcription at the general level of systems biology and networks is addressed. The next step is sequence-based studies of biologically linked genes and proteins that play a regulatory role in a variety of organisms. Finally, a detailed analysis including structure-based modeling of protein-DNA interactions at an atomic resolution is presented. At each level the characteristic challenges and results are described in detail. Furthermore, implications in disease-related research and the potential therapeutic aspects are discussed.

Systems biology aims at modeling whole networks, cells, tissues, or organisms [135, 273]. This clearly presents a true challenge, as most biological events are complex, multifactorial, highly interdependent, and not yet fully understood. It is necessary to assess the effects of a genetic alteration or a dysfunctioning transcriptional control signal at the whole-organism level [42]. A comparison of the gene expression in different cell or tissue types (e.g. normal and tumor) can reveal differentially expressed biologically related genes. Cancer genesis and HIV infection are two examples of diseases with an immense impact, to which no single therapy can be applied with success. Observations of the effects at the level of systems biology, in combination with more detailed studies of their individual causes and stages of progression, are inevitable for therapeutic [57, 115]. In Chapt. 4, a system for integrated analysis of cancer-related data (CAP) [67] is presented. It demonstrates the importance of bringing together heterogeneous data from different sources, in order to facilitate a full-picture analysis of potential differences in gene regulation [176]. Foremost, it provides evidence for the involvement of gene regulation in cancer development and it serves as a practical example of how putative drug targets and diagnostic markers for certain cancer types can be identified.

Complex regulatory networks underlie each phenotypic response that can be observed at the level of systems biology [136, 251]. The networks are highly complex, but the components i.e. transcription factors and the binding sites in their target genes can be evaluated at the sequence-based level. Advances in the area of bioinformatics during the last decades have contributed to the development of a broad range of analysis tools. Three sequence-based studies of the regulatory mechanisms controlling the expression of biologically linked sets of target genes are presented in Chapt. 5. These studies are similar in that they all focus on common mechanisms behind the transcriptional regulation of biologically-related genes, however, they differ in a number of fundamental ways. Several organisms and their genomes, *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae* [218], and *Arabidopsis thaliana* are analyzed, illustrating the different challenges each organism presents and the invaluable use of model organisms. Furthermore, the methodology by which the data sets containing the identified target genes were collected and the size of the data sets vary. A detailed study of the sequences of the participating genes constituting the individual building blocks of the network is necessary. This type of step-by-step analysis facilitates the reconstruction of the jigsaw puzzle any gene regulatory network presents [112].

Zooming in further, reaching molecular and atomic resolution, is a necessity for evaluating the thermodynamics and specificity of the interactions involved in protein-DNA complex formation [147, 175]. Macromolecular modeling and simulations enable studies of the flexibility and dynamics of proteins, DNA, and RNA. Inter- and intramolecular interactions important in protein-DNA recognition and transcriptional control can be studied in detail. The interactions in ZF-DNA complexes are the focus of the structure-based studies at atomic detail presented in Chapt. 6. The effects that base modifications exert on the binding free energy of a protein-DNA complex are explored using both explicit and implicit solvent models. This final part of the study serves an example of how theory and simulation hold considerable promise in analyzing the fundamentals of physical contributions of the binding process that eventually controls gene regulation [4].

The multitude of factors underlying transcriptional control is addressed at several different levels in this study. In particular, the sequence-specific recognition involved in gene regulation is addressed using integrated database annotation systems, sequence-based methods, and finally structure-based molecular simulation approaches. The individual levels constitute key sources of information. Sequence-based information is crucial for studying the components of a regulatory network, whereas structure-based studies can help explaining any redundancy observed in sequence-based models. Knitting the information together enables the levels to profit from each other. The results indicate that efficient, reliable, and sophisticated detection of functional binding sites ultimately relies on the intelligent integration of both experimental data and theoretical methods. Furthermore, the theoretical models complement experimental observations and add valuable insight into the thermodynamic properties of the binding process and which effects they exert on the cellular level. Continuous improvement is facilitated through feedback loops between experiments and theory. A detailed understanding of all aspects involved in protein-DNA recognition and binding serves as a cornerstone in disease-related research, rational drug design, and therapy.

An example resulting from successful communication and feedback between experiment and theory is the development of engineered TFs. The immense potential residing in the ability to control gene regulation at will clearly provides a powerful tool for many applications in biotechnology and disease therapy. Engineered TFs have potential use for correcting abnormalities in gene expression, e.g. as gene therapy for treating cancer [50], for inhibiting viral replication [115], and for controlling the differentiation of transplanted stem cells [16]. The zinc finger (ZF) motif is the most commonly used structural motif among eukaryotic TFs and has been observed to play key roles in the development of diseases like cancer [74, 234].

A ZF is a modular protein domain consisting of 30 amino acid residues that specifically recognize and bind to short subsequences (3-4 nucleotides) of DNA. The specific recognition of ZF TF binding sites has been studied in great detail, generating a vast amount of experimentally derived sequence data [249], affinity measurements [94, 132], and structural information [143, 199]. The ZF domain is a suitable framework for *de novo* engineering of DNA specificity [9, 60, 118]. Designing protein-based drugs for gene therapy have several advantages compared to other chemical approaches, especially when it comes to drug transport, combinatorial testing, synthesis costs, delivery, and side-effects [115, 159]. Furthermore, ZF-TFs can regulate endogenous genes, which competing methods such as antisense or RNA interference (RNAi) can not, and the expression can be either induced or repressed [151]. A very promising approach is to fuse engineered zinc finger domains with specific sequence recognition to a receptor domain that can be activated chemically e.g. using hormones [20]. ZF proteins have successfully been used for promoting angiogenesis by inducing the VGEF (Vascular Endothelial Growth Factor) TF [203]. Using biological molecules (such as proteins, peptides, DNA, and RNA) as lead compounds offer an attractive alternative to chemical (non-biological) compounds, since toxic side-effects can be avoided.

In summary, this thesis demonstrates different methods and their combination for analyzing gene regulation computationally. The results obtained at each individual level can profit from information obtained at other levels. All components are required, as the ambition is to create a complete understanding of gene regulation in order to enhance assessment of applications in biotechnology and therapeutics.

2 Theoretical background

Organism development, cellular responses to external factors, and maintenance are all crucial properties for survival [246]. All cellular processes are controlled in a meticulous and exact manner, which is the astonishing result of millions of years of evolution [5, 272]. Gene regulatory networks are dependent on a complex interplay between molecules, where spatial and temporal control of molecular interactions are of highest importance [83, 236].

The blueprint of life is encoded in the DNA (deoxyribonucleic acid), which is a fascinating molecule in its simplicity, as it is built up from four basic building blocks (called nucleotides)). DNA is constructed by two intertwined polynucleotide chains, between which complementary base pairs (bp) are formed. The characteristic right-handed helix structure was discovered in pioneering experimental studies by Franklin [79] and Wilkins [277]. The original diffraction pattern from 1953 is shown in Fig. 2.1. Later the same year, Watson and Crick [270] proposed a theoretical model of the DNA helix, which can be seen in Fig. 2.2. The DNA defines heredity and encodes all information needed for life, however, a DNA molecule alone is inactive (exception: catalytic nucleic acids). The functional biomolecules (actors) are the proteins, which are capable of interacting with DNA and participate in all cellular processes. The *central dogma* forms the backbone of molecular biology by providing a description of how proteins are generated from DNA and is usually described in four steps: replication, transcription, processing, and translation [37].

Essential to gene transcription are the specific interactions between DNA and proteins. Hence, the following sections start with a presentation of the concepts crucial to understanding the biological background, including DNA-binding proteins and their role in gene regulation. The biochemical properties of proteins and DNA are introduced, in order to facilitate the understanding of the chemistry and physics involved in the protein-DNA recognition and binding process. Finally, the concepts of computational chemistry that have enabled theoretical modeling of biomolecules, computational analysis of binding affinities, and monitoring of thermodynamics of complexes are described.

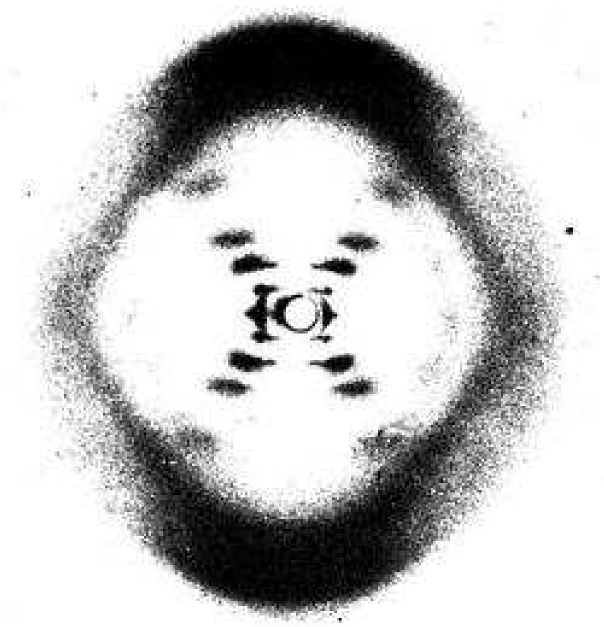


Figure 2.1: The original X-ray diffraction pattern by Franklin [79], which provided a basis for the theoretical description of the DNA helix formulated by Watson and Crick in 1953.

2.1 DNA-binding proteins

The main function of DNA is to store, replicate, and enable propagation of information (heredity) throughout evolution [37, 259]. Proteins, on the other hand, actively carry out functions in all cellular processes including enzymatic processes, signal transduction, transport, translation, and metabolism, specific through protein-protein, protein-ligand, and protein-DNA interactions [247]. Storage, packing, repair, protection, and transcriptional regulation of genes are important tasks, which all are mediated through protein-DNA interactions [56, 198]. The mechanisms underlying transcriptional control of gene expression and the proteins responsible for this control are the focus of the following two subsections.

2.1.1 Gene regulation

A gene was first described as a discrete unit of heredity that influences a visible trait. Later this description was extended to include that a gene contains the directions for making a protein and that its expression is controlled, which is a well-known fact today. Eukaryotes and prokaryotes differ in many aspects, e.g. genome size, number of genes, and in the complexity of gene regulation [167, 246]. Eukaryotic gene regulation is the focus of this study, however, the biochemistry behind specific protein-DNA interactions similar for both eukaryotes and

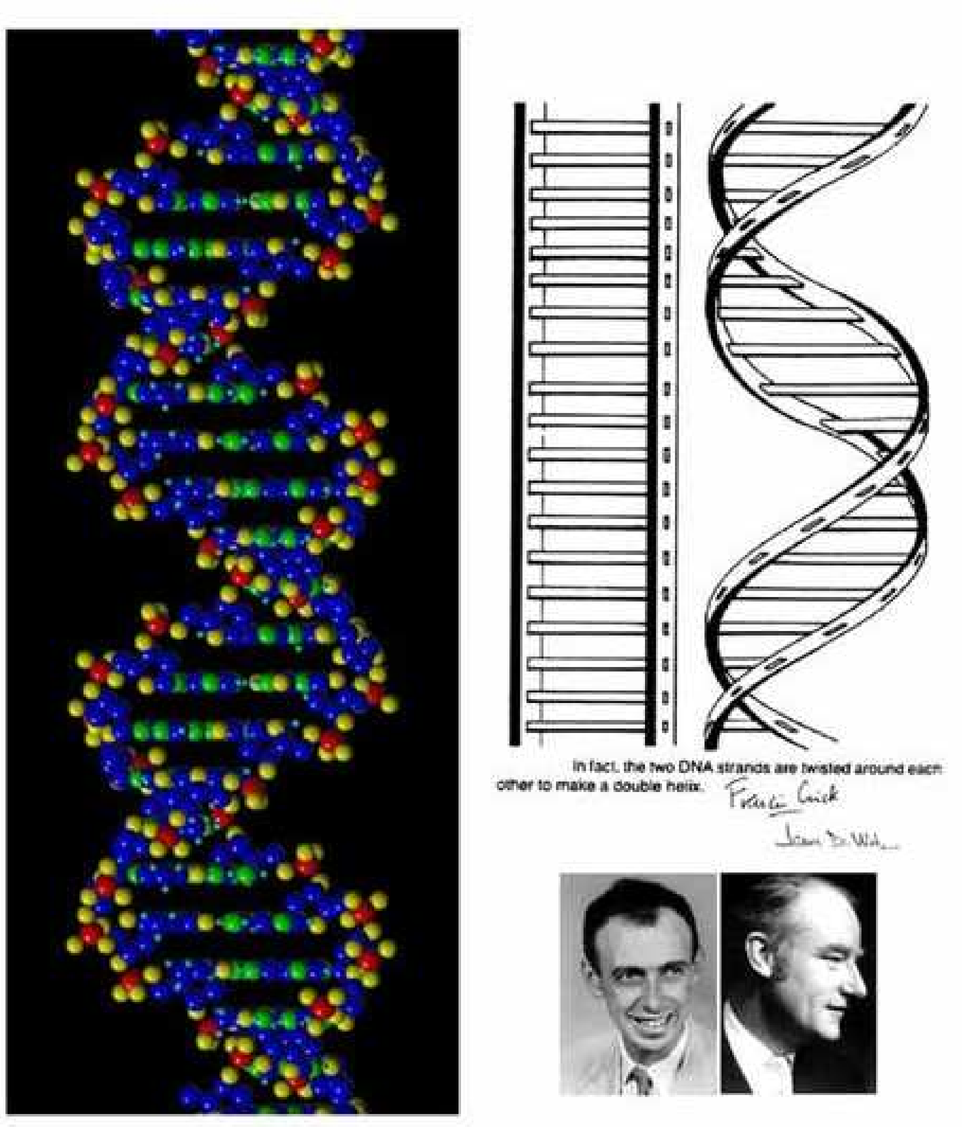

In fact, the two DNA strands are twisted around each other to make a double helix.
Francis Crick

Figure 2.2: The right-handed double-stranded DNA helix with its characteristic major and minor grooves, as described by Watson and Crick in 1953 [270].

prokaryotes.

The human genome contains roughly 25.000 genes [194, 263], which are all encoded in the genomic DNA. Most genes are split up into segments (exons), which are interspersed with segments of non-coding DNA (introns). Regulation of gene expression is a multistep process that occurs at several different levels in the cell: unwinding inactive DNA, transcription, mRNA splicing, export out of the nucleus, degradation and protection of mRNA, translation, and post-translational modification [37, 259]. Chromatin remodeling and modification are two still not-so-well-understood processes important in gene regulation [56].

Transcriptional regulation is carried out by a set of DNA-binding proteins, called transcription factors (TFs). The protein-DNA interactions are highly specific [264] and enable the TFs to recognize and bind to short transcription factor binding sites (TFBSs), typically located near the genes in non-coding genomic DNA. The non-coding DNA was first dismissed as non-functional, however, it became clear already at an early stage that these regions are involved in and responsible for regulatory processes by activating or silencing the surrounding genes [68, 187].

The regions where TFBSs are located are referred to as regulatory regions [81, 267]. Specifically, the closest (to the gene) region on the DNA is the proximal promoter and the more distant is a *cis*-regulatory element [43, 82, 30, 275]. The main differences between human and yeast promoters are graphically illustrated in Fig. 2.3. A set of basal TFs are responsible for recruiting and binding the transcription initiation complex (TIC), also enabling the RNA polymerase to bind to the promoter [8]. The TIC binds to the promoter at the location of the TATA box (shown in Fig. 2.3). The *cis*-regulatory regions contain a more diverse set of TFBSs and constitute the required key to the expression pattern of the affected genes [30]. The TF-*cis*-regulatory complex can be located several kilobases (kb) away from the transcription start site (TSS), however, the flexibility of the DNA molecule enables protein-protein interactions between these two TF complexes, thereby directing transcription [27, 82, 259].

A huge leap forward has been taken from the early studies of single genes to the large-scale analysis technologies that are available today. The efficient sequencing of the genomes from various organisms [263] has contributed to high-throughput analysis of gene expression profiles and functional analysis of proteins, which present an ever increasing body of available data. A good starting point is to study less complex (from a transcriptional point of view) model organisms, and then to transfer knowledge gained to a more complicated system such as human [112, 136, 218].

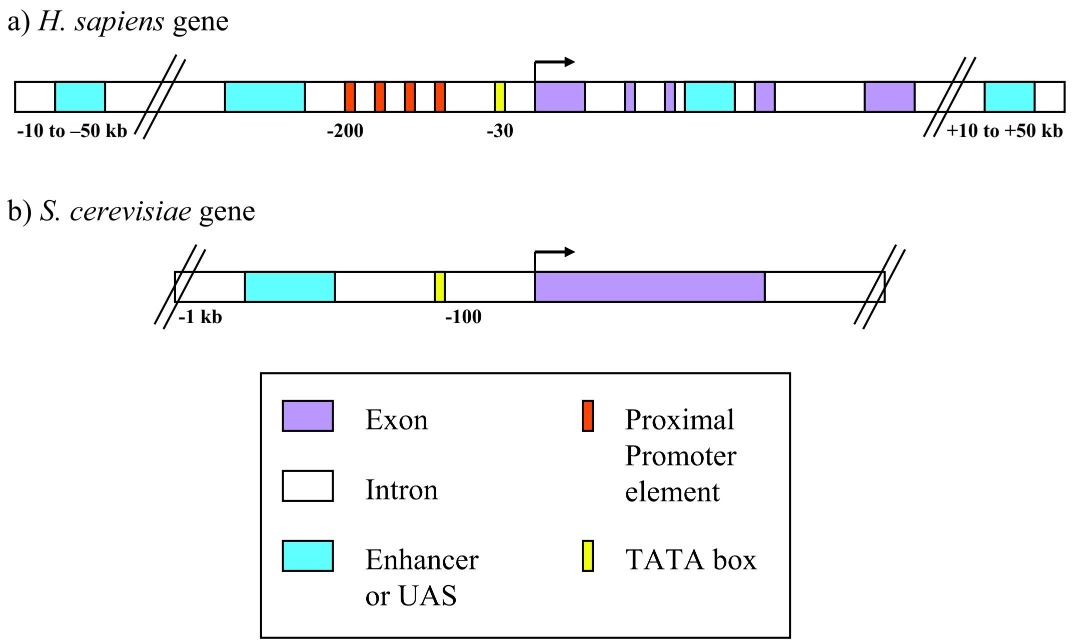


Figure 2.3: An illustration of typical regulatory elements and the differences (length and complexity) between a) *H. sapiens* and b) *S. cerevisiae* promoters and genes. The transcription start site (TSS) is indicated by an arrow.

Computational methods for analyzing and visualizing experimental data enable studies of most aspects of gene regulation. Tools exist for identifying genes, promoters, and *cis*-regulatory elements in DNA [28, 43, 82, 153, 157, 275]. Furthermore, algorithms have been developed for identifying co-expressed genes in gene expression data [88]. Eukaryotic transcription is a complex network of control mechanisms, where combinatorial control, synergy, activation, repression, feedback loops, and feed-forward loops are familiar characteristics [163, 167]. The complexity of the regulatory network controlling transcription in an organism is a good measure of the complexity of the corresponding organism [251]. Furthermore, it has been observed that many more TFs are involved in controlling the expression of one eukaryotic gene (compared to a prokaryotic gene), which implies that a more complex regulatory network is required in eukaryotes [53]. The regulation is context-dependent and the observed complexity ensures that an individual gene or groups of genes are expressed in the right cell type at the right time [75]. This spatial and temporal control of gene expression allows the organism to respond to external signals in an efficient way [268].

2.1.2 Transcription factors

The two main types of TFs are (i) TFs responsible for sequence specific recognition of the TFBSs within the *cis*-regulatory regions, and (ii) the set of basal TFs, except RNA polymerase itself, that bind to the proximal promoter and constitute the TIC. The underlying biochemical principles of protein-DNA binding are the same, however, this work is focused on the first type of TFs.

TFs typically have two functional domains, a DNA binding domain (DBD) and an activation domain. The DBD can recognize a specific DNA sequence and thereby bind to its target TFBS with high affinity [53]. The activation domain, on the other hand, is typically responsible for making protein-protein interactions with other surrounding TFs that also bind to the *cis*-regulatory element or to the TIC.

The TFBSs in the *cis*-regulatory regions are short DNA sequences, typically 5 to 12 bps long. Often several TFBSs are clustered within a *cis*-regulatory region and form functional modules [122, 267]. This enables a defined set of TFs to come together in a specific orientation and act in concert [30]. The combinatorial logic ensures a defined set of genes, e.g. all required for a cellular process, to be expressed simultaneously. Furthermore, these functional regions of non-coding DNA have been conserved throughout evolution, hence detection by sequence alignment is usually helpful [28, 68].

Estimates have proposed that about 6-7% of all eukaryotic proteins are TFs [36]. Classification of TFs into structural families is usually done according to the structural properties of the DBD [165, 190]. Fig. 2.4 illustrates three of the most common classes A) ZF (zinc finger), B) HTH (Helix-Turn-Helix), and C) bZip (basic leucine zipper).

Studies indicate that about 20% of all eukaryotic TFs belong to the ZF class [51]. ZF proteins acquire DNA-binding ability by Zn(II) complexation. In the ZF domain of the Cys_2His_2-type, each finger is about 30 amino acids long and forms a basic $\beta\beta\alpha$-fold. The Zn ion is tetrahedrally coordinated by the two Cys and His residues, as illustrated in Fig. 2.5. The side chains of the α-helix in positions -1, 2, 3, and 6 (with respect to the start of the α-helix) contact four DNA bps mainly on one strand in the major groove of the TFBS [71, 192, 280], as shown in Fig. 2.6. The ZF domain is highly modular and can be linked to other ZF domains using a linker peptide. The individual subsites overlap with one bp. The ZF-TF Zif268 (or early growth factor (EGR1), PDB code: 1AAY [71]) consists of three ZF domains and recognizes a 10 bp consensus sequence. The primary contacts between the side chains of Zif268 and the DNA bases are illustrated in Fig. 2.6. The DNA major groove is slightly widened, however, the DNA helix is not significantly bent upon binding.

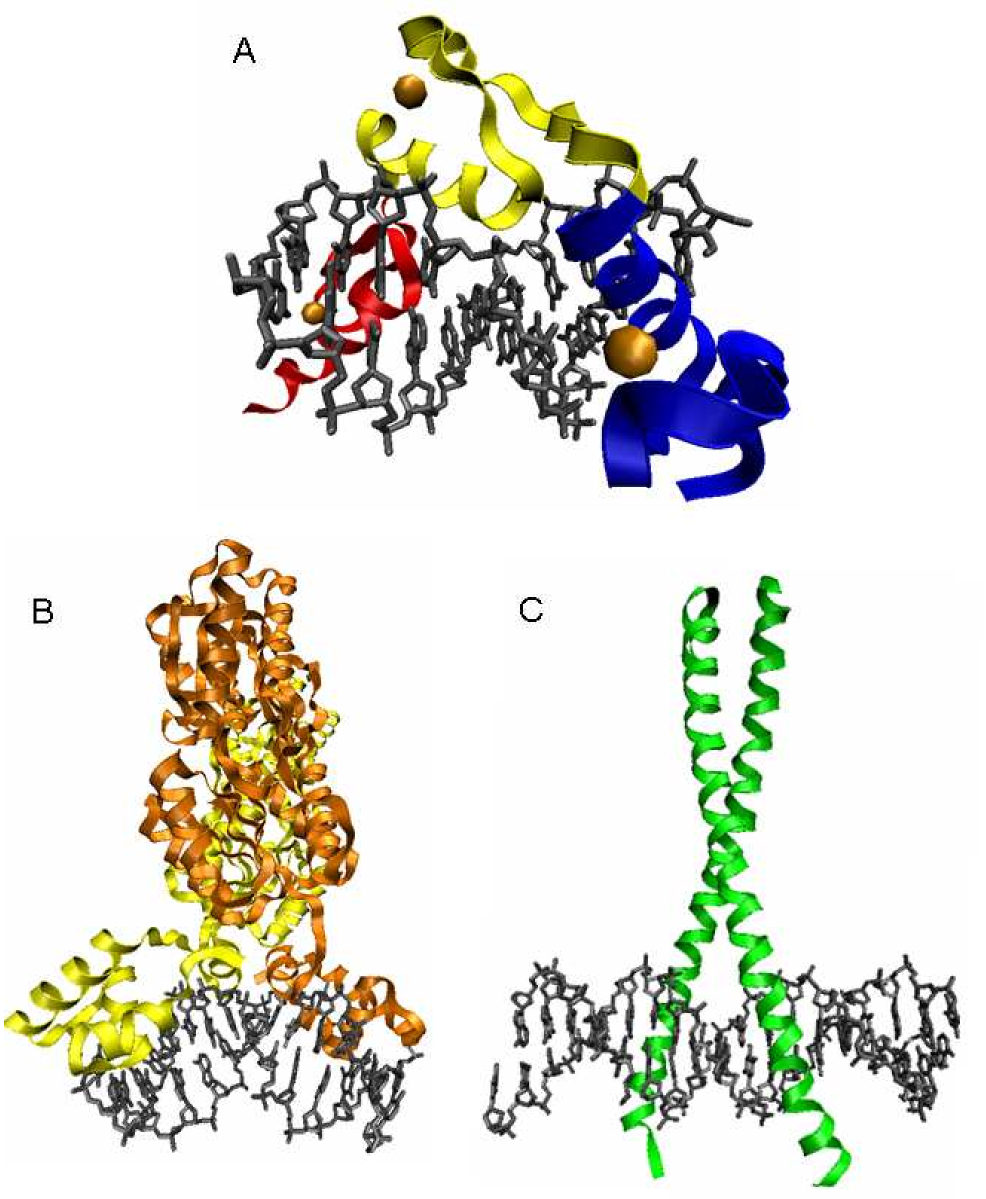


Figure 2.4: There are several structural families among DBDs in TFs. Three of the most common ones are the: A) Zinc Finger (ZF) family, which is illustrated by the three ZF domains of the Zif268 TF (PDB code: 1AAY [71]) that are linked to each other like train cars. Each ZF domain (here illustrated with different colors: red, yellow, and blue) recognizes a four bp long subsite (one bp overlap) of the 10 bp consensus sequence of the DNA (grey). The zinc ions are colored orange. B) The helix-turn-helix (HTH) motif is found in the DBD of the Lac repressor protein (PDB code: 1JWL [21]), which binds to two half sites of DNA. C) The yeast TF GCN4 (PDB code: 1YSA [70]) has a basic leucine zipper (bZip) motif (colored green).

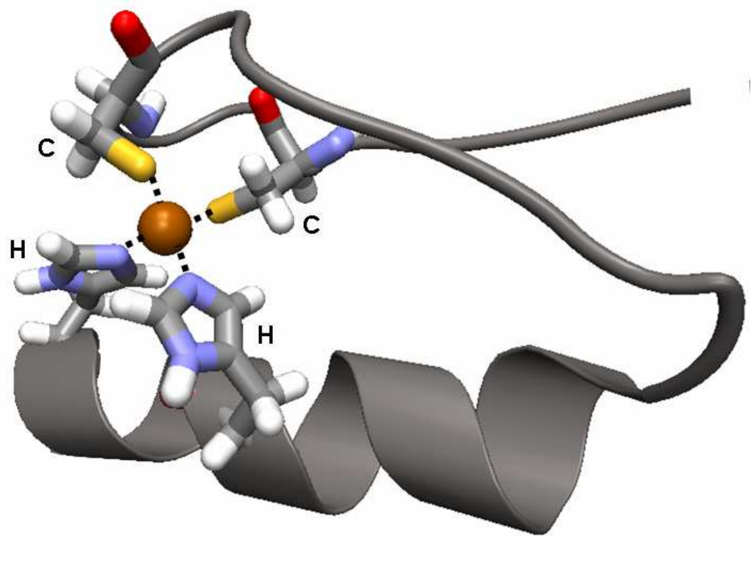


Figure 2.5: The ZF motif (grey) is one of the most common structural motifs in DNA-binding proteins. In the Cys_2His_2 ZF-class, two Cys side chains (C) from the β-sheets and two His side chains (H) from the α-helix tetrahedrally coordinate the zinc ion (orange).

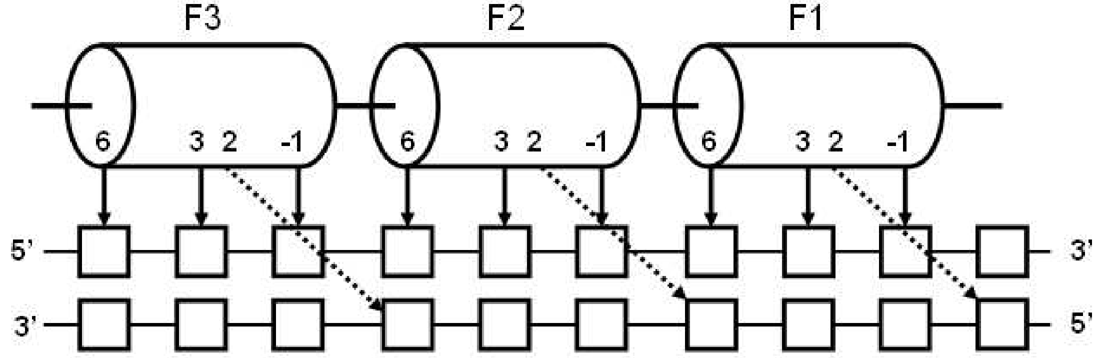


Figure 2.6: The primary contacts and the binding mode of the ZF-TF Zif268 are illustrated here. The recognition helices of the ZF motif are represented as cylinders. The contacts between the protein side chains (at positions -1, 2, 3, and 6 with respect to the start of the α-helix) and the DNA bases, are illustrated as arrows.

2.2 Protein-DNA recognition

Protein-DNA binding is a reversible process and the binding affinity (strength) varies greatly between different complexes. This facilitates induction, repression, and fine-tuning of gene expression. The affinity of the non-covalent association depends on the structural and biochemical complementarity of the two interacting molecules.

A general recognition code systematically describing the interaction preferences between protein side chains and DNA nucleotides, would be of great use in rational design of artificial TFs. The existence of such a code was hypothesized at an early stage [137, 188, 226]. However, the idea was dismissed as the first set of experimental protein-DNA structures had been solved and no simple one-to-one correspondence could describe the observed interactions [49, 173, 248]. Several studies, which address the specific recognition and aim to construct models of the interaction, using the framework of ZF-TFs (illustrated in Fig. 2.6) have followed [25, 40, 59, 189, 279]. It can be concluded that specific preferences between amino acid side chains and nucleotides do exist, however, some of the defined preferences are degenerate and are applicable to only one structural class of protein-DNA complexes [24, 165, 190].

The main physical and biochemical aspects that are important for understanding how specific protein-DNA binding is facilitated are introduced in the following sections. The factors governing and opposing complex formation are discussed in terms of thermodynamics and types of interactions that can be found in protein-DNA complexes.

2.2.1 Physical properties of proteins and DNA

The nucleotides are the monomeric units of the helical DNA polymer (which is about 20 Ångström (Å) in diameter). Each nucleotide unit consists of a pentose (deoxyribose), to which a phosphate group and a nitrogen-containing base are attached. The negatively charged DNA backbone is an alternating sugar-phosphate sequence, where the deoxyribose sugars are joined at both the 3'- and 5'-hydroxyl groups to phosphate groups. An illustration of the DNA backbone is provided in Fig. 2.7. The DNA chain elongation takes place at the 3' end of the DNA strand, where new phosphodiester bonds are formed. There are five different types of bases: adenine (A), guanine (G), cytosine (C), thymine (T), and uracil (U) displayed in Fig. 2.8. A and G are the larger purine bases, whereas C and T are the smaller pyrimidine bases. U is an additional pyrimidine base, which occurs in RNA (ribonucleic acid) but not in DNA. U is similar to T but has a single hydrogen at the position of the methyl group in

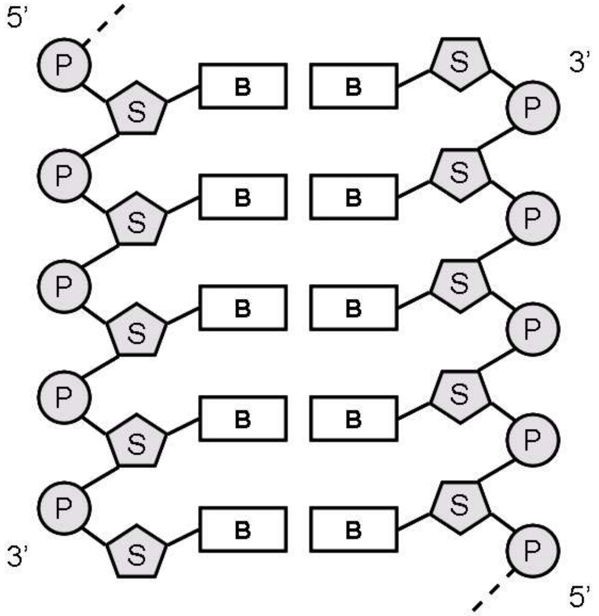


Figure 2.7: The sugar-phosphate (denoted S and P respectively) backbone is negatively charged. The complementary bases (B) form a ladder-like structure.

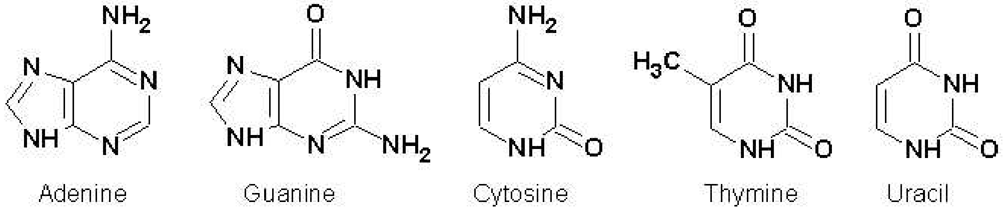


Figure 2.8: The five different bases Adenine (A), guanine (G), cytosine (C), thymine (T), and uracil (U). Possible base pairs are A-T and G-C in DNA, whereas A-U and G-C occur in RNA.

the base T.

The bases are hydrophobic and form complementary bps through hydrogen bonds, which are shown as dotted lines in Fig. 2.9. Possible bps are A-T and G-C (additionally, A-U in RNA). The edges of the bps form unique patterns of hydrogen bond donors and acceptors in the DNA major groove (facing upwards in Fig. 2.9) [129]. The profile of the edges of the bases is unique for all possible bps (A-T, T-A, G-C, and C-G) in the major groove, whereas in the minor groove, it is only possible to distinguish between A-T/T-A and G-C/C-G bps. The major groove is wider (12 vs. 6 Å) and deeper (8.5 vs. 7.5 Å) than the minor groove. Most TFs bind to the major groove, as it is wider and presents a unique pattern for specificity, but some make contacts with both grooves in order to increase the overall stability of the complex.

Proteins are polypeptides - a chain of amino acid residues connected by peptide bonds.

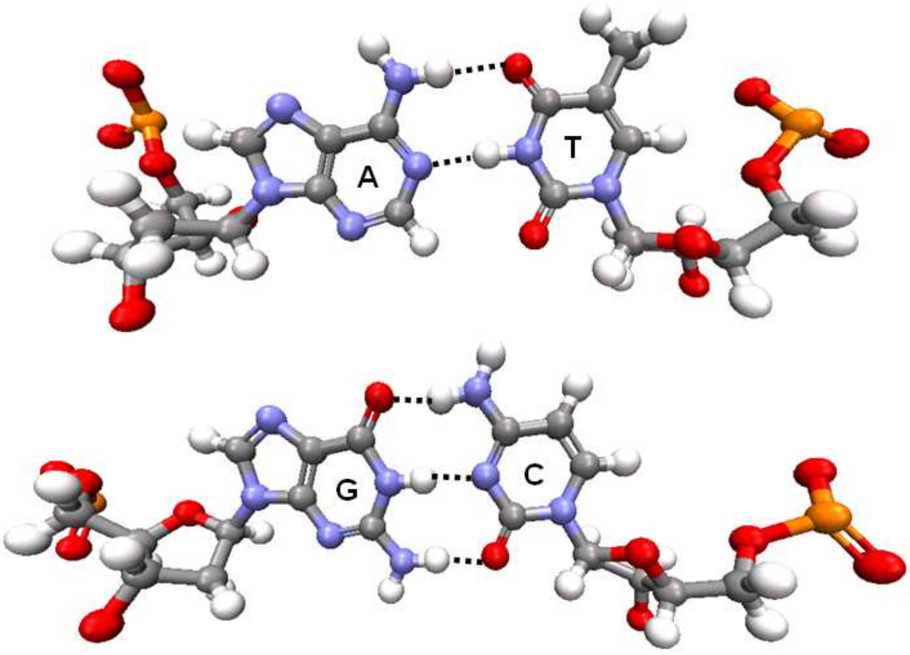


Figure 2.9: Intermolecular hydrogen bonds (dotted lines) in the bps stabilize the DNA double helix. The edges of the bps form a pattern of hydrogen bond acceptors and donors that can be recognized by amino acid side chains of TFs. The pattern is unique for each bp (A-T, T-A, G-C, and C-G) in the major groove (up), whereas it is only possible to distinguish an A-T bp (top) form an G-C bp (bottom) in the minor groove (down) [9]. The sugar-phosphate backbone is independent of the bp sequence.

The side chains of the amino acids differ in length and physical properties, such as charge and hydrophobicity. Under biological conditions (in water and at physiological pH and temperature) proteins fold and form structural motifs. The DBD in TFs is the structural motif that recognizes and binds to DNA [190], as illustrated in Fig. 2.4. The α-helix is a common structural motif in DBDs [165], from which the side chains responsible for DNA recognition can extend into the DNA grooves. The negatively charged sugar-phosphate DNA backbone is independent of the nucleotide sequence and attracts positively charged side chains (such as Arg and Lys) of the TFs.

Structural adaptation and biochemical complementarity are two features crucial to complex formation, which are achieved through conformational changes of both molecules [63, 221, 255]. The DNA is flexible and often distorted when bound to a TF [62], which is illustrated by the two examples displayed in Fig. 2.10. The Catabolite gene Activator Protein (CAP) distorts DNA to a 90° bend, whereas the tandem Zif268 TF (six-subunits) wraps the DNA without causing helix distortion. An induced conformational change, resulting in bending or twisting of the helix, can expose the necessary contact points for high affinity binding. G-C rich regions in genomic DNA are less flexible than A-T rich regions, since G-C bps contain 3 hydrogen bonds (one more than A-T bps, as shown in Fig. 2.9).

2.2.2 Specificity through protein-DNA interactions

Sequence specificity is necessary for recognition of TFBSs in genomic DNA sequences. The biochemical properties of the DNA backbone are independent of the base sequence, whereas the bp edges and the helix flexibility are dependent on the nucleotide sequence. Structural and thermodynamic studies of various protein-DNA complexes have contributed to a rich source of information on protein-DNA interactions. The interactions are either *direct* or *indirect*, which both contribute to the protein-DNA affinity [95, 227, 240].

The side chains of the protein and the nucleotides of the DNA molecule allow for a wide range of *direct* interaction possibilities. Electrostatics play an important role in the direct interaction governing protein-DNA complex formation, as the biomolecules are charged and surrounded by water. The direct interactions can be obtained by analyzing structural data for protein-DNA complexes and are used for creating statistical pair-wise interaction potentials [141, 168].

The direct contacts can be further classified as *specific* or *non-specific*. The specific interactions enable TFs to recognize and bind to certain target TFBSs in genomic DNA. These

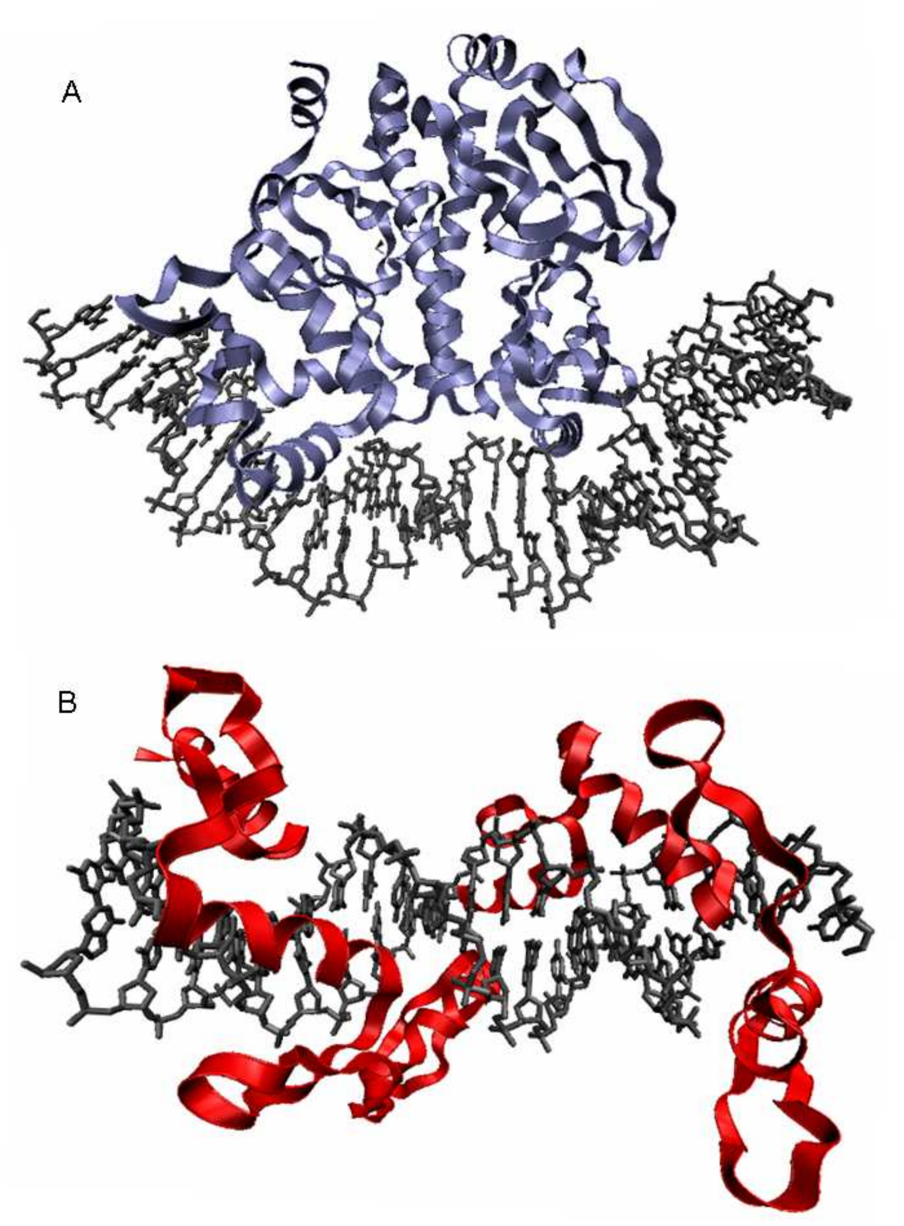


Figure 2.10: Conformational changes of either the DNA helix or the TF are common upon complex formation. A) The Catabolite gene Activator Protein (CAP) shown in blue, PDB code: 1CGP) bends DNA (grey) with 90°. B) A tandem Zif268 on the other hand (a six-subunit ZF protein shown in red, PDB code: 1P47), wraps around the DNA helix without bending it.

are either formed between the edges of the DNA bps and the amino acid side chains, or between the bp edges and the protein backbone. Non-specific interactions, on the other hand, are independent of the DNA sequence and are formed between the DNA backbone and the protein backbone or the DNA backbone and the amino acid side chains.

Hydrogen bonding is an example of intermolecular interactions that contribute directly to the binding specificity. The unique pattern of hydrogen bond donors and acceptors on the bp edges in the major groove form anchor points for intramolecular hydrogen bonds (described in Fig. 2.9). The hydrogen bonds can be of different type: *single*, *bidentate*, and *complex* [166]. A single hydrogen bond is formed between one atom in the protein and one atom in the base. A hydrogen bond of the bidentate type is formed if one atom of the protein interacts simultaneously with two atoms of one base, or if two atoms of the protein interact with two atoms of one base. Finally, a complex hydrogen bond is formed if two or more hydrogen bonds are formed between two (or more) atoms of the protein and two or more nucleotides in the DNA molecule.

There are also water-mediated hydrogen bonds that facilitate additional specific long-range interactions, which are crucial in the recognition process and for forming complex networks at the interaction surface [119]. Structure-based studies have shown a relationship between resident water molecules and important protein-DNA interaction points [278].

Hydrophobic interactions are highly important, since all interactions in the cells take place between molecules that are solvated by water. Complex formation is promoted by shielding hydrophobic residues, hence these residues are commonly found in the interior of folded proteins or at the binding interfaces of biomolecular complexes.

The positively charged protein side chains Arg, Lys, and His are common in DBDs, since they can form ionic bonds with the negatively charged DNA sugar-phosphate backbone.

The *indirect* contributions to the binding energy are far less understood and present a challenging problem to analyze and evaluate [72]. These include the influence of conformational changes on the internal energies of the two molecules, desolvation, distortion, and further macromolecular changes that occur upon complex formation [241, 240]. It is essential to regard all different types of interactions, since it has been shown that no individual interaction alone can explain the binding specificity [123, 166, 169]. In many cases, not even combinations of the different types are sufficient [285]. Hence, the contribution of the indirect interactions can not be neglected and remains to be explored [59, 189, 279]. Open questions include the interdependence between neighboring bps in the DNA and amino acids in the protein [22, 39], synergy between neighboring DBDs (e.g. ZF units) [47, 113, 180, 138], flexibility of the protein and the DNA helix [93, 177, 191], hydration at the interface [63, 204, 225, 256, 278], and the role of counterions [178, 202, 262].

2.2.3 Thermodynamics of binding

Protein-DNA binding is a multistep process [264], in which the two molecules first get together by diffusion. Subsequent binding occurs if they have chemical and structural complementarity at their interaction surfaces. The binding process is directly dependent on several factors: for example the concentration of the two interacting molecules, solvation, temperature, and the concentration of counterions in the solution [204].

The Gibbs free energy (G) of a system is dependent on the internal energy (enthalpy H), the internal microscopic disorder (entropy S), and the temperature (T) [148]. The change in Gibbs free energy is a thermodynamic quantity, which is defined by

$$\Delta G = \Delta H - T\Delta S.$$

The relative enthalpic and entropic contributions determine the sign of the ΔG (a negative sign indicates a spontaneous process), eventually determining if a certain protein and a DNA sequence form a complex. Ionic (charged) interactions, desolvation effects (polar and non-polar interactions), hydrogen bonds (direct and water-mediated), and van der Waals forces belong to the enthalpic contributions. The entropic contributions to ΔG originate from desolvation effects, losses in degrees of freedom of side-chains, and losses in degrees of freedom due to limited molecular flexibility i.e. translational and rotational motion relative to the binding partner (these are typically constant).

Gibbs free energy determines how much work is attainable for a given process and is defined by

$$\Delta G = -RT \cdot lnK_d$$

where R is the gas constant and T the absolute temperature (measured in Kelvin). ΔG is measured in kJ/mol (SI units), however, often given in kcal/mol. The dissociation constant K_d is a measure of the binding strength between the two interacting molecules and is defined by

$$K_d = \tfrac{1}{K_a} = \tfrac{k_2}{k_1} = \tfrac{[A][B]}{[AB]}$$

where k_1 is the rate of dissociation, and k_2 is the rate of association between the two molecules A and B. K_d is inversely proportional to the association constant K_a. If a high concentration of the molecules A and B is required for complexation, the strength of the binding is low. TFs typically have a K_d in the range from 10^{-8} to 10^{-11} (nM range).

The relative binding free energy ($\Delta\Delta G$) between two complexes with different affinities (ΔG_1 and ΔG_2) is defined as

$$\Delta\Delta G = \Delta G_2 - \Delta G_1 = -RT \cdot ln\frac{K_{d2}}{K_{d1}}$$

The ΔG for a complex (thereby also the $\Delta\Delta G$) is dependent on several types of interactions, which will be discussed in detail in the following section.

2.2.4 Types of interactions

The interactions governing and opposing binding vary in strength and abundance between different protein-DNA complexes [189]. *Covalent* bonds are strong bonds (50 to 110 kcal/mol) and occur when one or more electrons are shared between two intramolecular atoms (e.g. the peptide bonds in the protein backbone). The *non-covalent* bonds are weaker (1 to 5 kcal/mol) than covalent bonds and therefore the interaction is easily reversible. These bonds occur both intermolecularly (between two molecules forming a complex, e.g. in protein-DNA and protein-protein complexes) and intramolecularly (within a protein e.g. secondary structures). The non-covalent interactions differ in their nature, strength, number, and optimum distance [148]. The most important ones for complex formation are the electrostatic interactions and hydrophobic interactions. Electrostatic interactions play a role in all charged interactions, which include ionic bonds, hydrogen bonds, and van der Waals interactions. Affinity high enough for complexation is achieved if the total sum of different types of interactions favors binding. Note that all weak bonds typically have an effect at short distances of a few Ångström only.

Electrostatic interactions (E_{ES}) arise between two charges (i and j) and are described by Coulomb's law. Equal charges repel, whereas opposite charges attract. The electrostatic interactions are proportional to the product of the electrical charges (q_i and q_j) and inversely proportional to the distance between the charges (r_{ij}) and the permittivity of vacuum (ε_0).

$$E_{ES} = \frac{1}{4\pi\varepsilon_0} \cdot \frac{q_i q_j}{r_{ij}}$$

A partial charge is formed due to an uneven charge distribution inducing partially positively and negatively charged areas around the atom. Ionic bonds, hydrogen bonds, and van der Waals interactions are all electrostatic interactions, which are important for biomolecules solvated in water.

Ionic bonds are among the strongest non-covalent bonds and one bond can account for up to 5 kcal/mol of the binding energy. An ionic bond (salt bridge) is formed by favorable electrostatic interactions between two oppositely charged ions. At pH 7 the acidic (Asp and Glu) and basic amino acid residues (Arg and Lys) have ionized side chains. An example of a salt bridge is illustrated in Fig. 2.11 (A). Carboxyl groups have lost a proton and carry a charge of -1 (which is delocalized over the two oxygen atoms), while amino groups have gained a proton and carry a charge of +1 (which is delocalized over the three hydrogen atoms). Ionic bonds (red dotted line) are typically about 2.7-3.0 Å.

Hydrogen bonds are almost as strong as ionic bonds and can account for up to 5 kcal/mol per bond. A hydrogen bond is formed between a donor and an acceptor group. The donor group contains a ploarized hydrogen, which it then shares with the acceptor group. To facilitate this the donor group is a strongly electronegative heteroatom (such as O, N, and F). The electron cloud of the hydrogen is decentralized towards the electronegative hydrogen bond *donor* leaving the hydrogen with a positive partial charge. This positive charge can attract the free electron pair of another heteroatom, which becomes the hydrogen bond *acceptor*, as illustrated in Fig. 2.11 (B). The strength of the hydrogen bond is highly dependent on the distance and the angles between the involved atoms. The optimum distance between the hydrogen bond donor and the acceptor in water is 2.8-3.2 Å, which is approximately 2.0 Å between the hydrogen and the acceptor (red dotted line in Fig. 2.11 (B)). The optimal geometry is achieved if the angle α is 150-180°, and the angle β is 100-180° (blue dotted lines). A special type of hydrogen bond is the **water-mediated hydrogen bond**. If the hydrogen bond donor and acceptor lie to far apart to interact directly, a water molecule (H_2O) can bridge the distance between them. Water-mediated hydrogen bonds can cause some water molecules to become “trapped” between the molecules at the interaction interfaces [278].

van der Waals interactions occur between temporarily induced dipoles carrying a free electron pair. These interactions are also called electrodynamic interactions, as dipole oscillations occur within the interacting molecules. The van der Waals forces are typi-

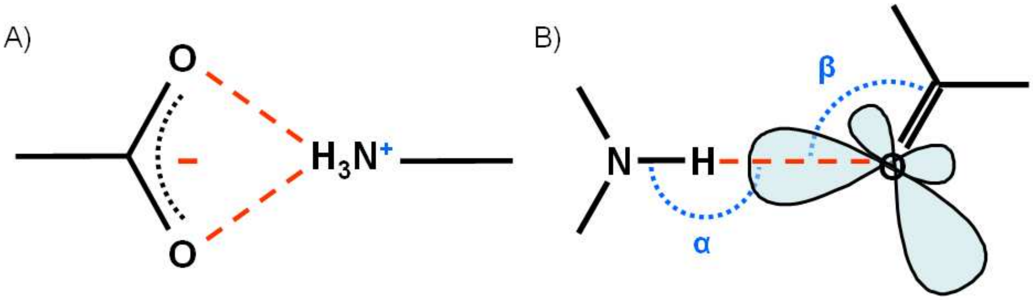


Figure 2.11: Favorable electrostatic interactions lead to the formation of ionic bonds between charged groups. The carboxyl group is negatively charged at pH 7, whereas the amino group is positively charged (shown in A). A hydrogen bond is formed between a hydrogen attached to a highly electronegative hydrogen bond donor atom, thus carrying a partial positive charge, and the free electron pair of a hydrogen bond acceptor, as shown in B. The optimal geometry is important to the strength of the hydrogen bond (details described in the text).

cally weaker than both ionic and hydrogen bonds, contributing with about 1 kcal/mol to the binding energy.

Hydrophobic interactions are slightly weaker than hydrogen bonds. Water molecules are polar and form an ordered pattern of interactions (a network of hydrogen bonds), whereas non-polar molecules are incapable of doing this. Therefore, shielding non-polar patches of molecules from the surrounding water will contribute positively to complex formation by decreasing the entropy of the system.

2.2.5 Binding affinity

The binding free energy (ΔG) is proportional to the statistical probability that a certain TF binds to a specific DNA sequence. The probability can be obtained using the *partition function* (sum-over-states), which describes the statistical properties of a system. In classical statistical mechanics the partition function is expressed as an integral, since the microstates are uncountable. In the case of discrete DNA binding sites the partition function can be represented as a sum. If the microstates that the system can occupy are denoted x ($x = 1, 2, 3, ...$), as illustrated in Fig. 2.12, and the energy of a system in microstate x is denoted E_x, then the partition function Q is defined as

$$Q = \sum_x e^{-\beta E_x}$$

where

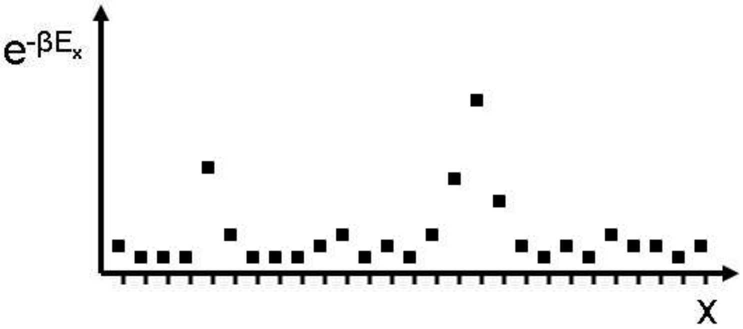


Figure 2.12: The partition function describes the statistical properties of a system. In the case of TFBSs in DNA, the partition function is described as the sum (Z) of $e^{-\beta E_x}$ (where E_x is the binding energy) over all discrete microstates (x).

$$\beta \equiv \frac{1}{k_B T}$$

and k_B is the Boltzmann constant.

The statistical representation provides a means to calculate thermodynamic properties of the system. The probability P_i to find a system in any given microstate i is given by

$$P_i = \frac{1}{Q} \cdot e^{-\beta E_i}$$

which is also known as the Boltzmann factor. The partition function Q is used for normalization (ensuring that the probabilities add up to 1) which gives the probability P_i at equilibrium.

The average binding energy ($< E >$) for a protein over all possible binding sites x is defined as

$$< E > = \sum_x P_x \cdot E_x = -\sum_x (P_x \cdot lnP_x) - lnQ$$

The exact probability distribution Q could be obtained for each TF if it were possible to measure the binding affinities to all possible binding sites i. However, for practical reasons this is usually approximated from a small set of experimental measurements. The relative binding free energy ($\Delta\Delta G$) is the difference between two probability distributions (U and W), which is defined as

$$E(U_x, W_x) \equiv -\sum_i P_x \cdot log\frac{U_x}{W_x}$$

If the two distributions are similar for the states with high probability, their relative entropy is close to zero. This formalization defines the relationship between the specificity and the interaction energy in protein-DNA recognition. A low interaction energy indicates a high probability that the binding site will be selected.

2.3 Computational chemistry

Important advances in the field of computational chemistry are the result of merging classical mechanics with current knowledge of physical properties of biomolecules [175]. The introduction of mathematical energy functions and recent increase in computational power enabled the first complex calculations and simulations of small molecules, which sequentially lead to simulations of more complex biomolecules. Theoretical models of molecular systems facilitate studies of structural and dynamical properties of biomolecules, and have laid the foundation for simulations of biological systems and phenomena.

The basic theory of molecular mechanics and dynamics is introduced in the following two subsections, however, the reader is referred to excellent literature [148, 206, 205, 221] for more detailed descriptions. *Computational alchemy*, a commonly adopted technique nowadays, is outlined in the third and final subsection. These concepts are required for Chapt. 6.

2.3.1 Molecular mechanics

There are a few essential ingredients to molecular mechanics theory, the description of a structure (atomic coordinates), their relations (connectivity), and an empirical energy function with its parameters defining the potential energy (E) of the molecule [148]. The Born-Oppenheimer approximation reduces the complexity significantly by separating the movement of the nucleus from that of the electrons [7]. This separation is possible since the mass of the nuclei is far greater than the mass of the electrons surrounding the nuclei, leading to highly different velocities. The Born-Oppenheimer approximation enables energy calculations of macromolecules with up to 10^5 atoms. Biomolecules are then approximated as a system of point masses (atoms) that are connected by springs (bonds).

The potential energy of a system is calculated as the sum of a set of individual energy terms. Several energy functions exist, a typical molecular mechanics energy function (such as the AMBER force field [266]) consists of the following terms: a bond length potential (E_{bond}), a bond angle potential (E_{ang}), a torsional potential (E_{tor}), a Lennard-Jones potential (E_{LJ}), and a Coulombic potential ($E_{Coulomb}$). Thus, the total potential energy (E) is defined as

$$E = E_{bond} + E_{ang} + E_{tor} + E_{LJ} + E_{Coulomb}$$

The three first terms of the force field contribute to the *bonded* energy, whereas the Lennard-Jones and the Coulombic terms define the *non-bonded polar* (van der Waals) and *non-bonded electrostatic* (ionic) energies, respectively. Each term is described in more detail in the following.

The bonds are represented using harmonic potentials ($E_{bond_{ij}}$), in which the displacement from the optimal bond length ($l_{ij} - l_{ij_0}$) and a force constant (k_{ij}) define the strength of the bond between to atoms i and j.

$$E_{bond_{ij}} = \sum_{bonds} \frac{k_{ij}}{2}(l_{ij} - l_{ij_0})^2$$

The bond angle is represented by a similar term ($E_{ang_{ijk}}$), in which the potential is set for the optimal angle (θ_{ijk_0}) between three atoms ijk using a force constant (k_{ijk}).

$$E_{ang_{ijk}} = \sum_{angles} \frac{k_{ijk}}{2}(\theta_{ijk} - \theta_{ijk_0})^2$$

A rotational barrier for a bond between two groups of atoms ($ijkl$) is represented using a torsional potential ($E_{tor_{ijkl}}$). An example of a periodic rotational barrier (where n is the periodicity) is the preference of the two methyl groups of ethane to stay in *staggered* rather than *eclipsed* conformation. The height of the barrier is defined by the constant V_{ijkl} and the angle is defined by ω_{ijkl}.

$$E_{tor} = \sum_{torsions} \frac{V_{ijkl}}{2}(1 + cos(n\omega_{ijkl} - \omega^{o}_{ijkl}))$$

The van der Waals interactions are represented using an attractive-repulsive Lennard-Jones potential (E_{LJ}), which models electrodynamic repulsion at short interatomic distances and attraction between two atoms i and j at long interatomic distances. The ε is the dielectric constant, σ_{ij} is the equilibrium interatomic distance, and r_{ij} the actual interatomic distance.

$$E_{LJ} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (4\varepsilon[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}])$$

Finally, the Coulombic potential ($E_{Coulomb}$) among all pairs of charged particles in the system, defines the energy between two point charges q_i and q_j, which is dependent on the distance r_{ij}.

$$E_{Coulomb} = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{q_i q_j}{r_{ij}}$$

2.3.2 Molecular dynamics

Molecular dynamics (MD) is Newton's Law of Motion ($\mathbf{F}_i = m_i\mathbf{a}_i$) applied on an atomic level [148]. A force $\mathbf{F}_i$ is exerted on each atom in the system. The force is a product of the mass (m) and acceleration ($\mathbf{a}_i$) of the atom. The system's motion is propagated through time and space according to the second order differential equation of Newton's Law of Motion:

$$\frac{d^2\mathbf{r}}{dt^2} = -\Delta\frac{V(\mathbf{r})}{m} = \frac{\mathbf{F}}{m} = \mathbf{a}$$

The equation has to be solved numerically, since no analytical solution exists for the general case. It is solved simultaneously for each atom in the system to obtain its new positions and velocities at finite time steps $t + \delta t$. For a more detailed explanation of MD and its applications, the reader is referred to other excellent literature sources [44, 148]. Being able to follow the relative positions of the atoms as a function of time has opened the doors for studying the dynamics and flexibility of a molecule (or of a complex of molecules).

2.3.3 Treatment of solvent

In order to obtain reliable theoretical simulations of biological processes in aqueous solution accurate methods for calculating solvation free energies are required [204, 225, 256, 278]. Charged and polar groups are present in all biological macromolecules and DNA is one of the strongest polyelectrolytes, with approximately one electron charge per 1.7 Å [170] of the backbone. Hence, the electrostatic interactions in protein-DNA complexes are very important in aqueous solution. The developed methods for simulating the properties of the solvent vary widely in complexity and the ease of calculation. The molecular modeling package AMBER 7 [44] allows for both explicit and implicit treatment of the solvent.

Explicit solvent

Explicit solvent models typically yield good agreement with experimental data. However, they are computationally demanding for macromolecules like proteins and DNA, due to the often many thousands of water molecules needed for representing the solvent around the solute. Effects such as solvent-solute hydrogen bonds and charge screening, are important biomolecular effects [287] that are taken into account by these models. A commonly used explicit solvent model is the three-centered water molecule e.g. the transferable interaction potential three-point model (TIP3P) [125].

Implicit solvent

Implicit solvent models offer attractive alternatives as they consider the bulk properties of the solvent as a continuum and allow for shorter computation times [91, 139, 243]. The free energy of solvation, ΔG_{solv}, is the free energy change required to transfer a molecule from vacuum into solvent medium. The ΔG_{solv} can be partitioned into electrostatic, van der Waals, and cavitation free energies between the molecule and the solvent.

$$\Delta G_{solv} = \Delta G_{elec} + \Delta G_{vdW} + \Delta G_{cav}$$

The electrostatic contribution to the solvation free energy (ΔG_{elec}) is important for biomolecular systems. In 1920, Born derived an expression for transferring an ion with a defined radius a and charge q from vacuum into solvent medium with the dielectric constant ε.

$$\Delta G_{elec} = -\tfrac{q^2}{2a}(1 - \tfrac{1}{\varepsilon})$$

This expression laid the basis for the Generalized Born (GB) models of the electrostatic contribution to the solvation free energy [148]. The GB model comprises a system of particles with Born radii a_i, charges q_i, and interparticle distances r_{ij}. The induced response (polarization) of the solvent is modeled using the dielectric constant ε. Still and co-workers [243] defined an approximate expression for the free energy of solvent polarization for an arbitrary charge distribution of N charges:

$$\Delta G_{elec} = -\tfrac{1}{8\pi}(\tfrac{1}{\varepsilon_0} - \tfrac{1}{\varepsilon})\sum_{i=1}^{N}\sum_{j=1}^{N}\tfrac{q_i q_j}{f_{GB}}$$

f_{GB} is a smoothing function dependent on the interatomic distances (r_{ij}) atomic radii (a_{ij}). The function is defined by $f_{GB} = \sqrt{r_{ij}^2 + a_{ij}^2 e^{-D}}$, where $a_{ij} = \sqrt{a_i a_j}$ and $D = r_{ij}^2/(2a_{ij})^2$.

Choosing the proper dielectric constant for various simulations is a controversial issue in literature [224]. For proteins and DNA dielectric constant ε is in the range of 2-4 and the solvent surrounding the solute has a relatively high dielectric constant of approximately 80 [78].

The non-electrostatic contribution ($\Delta G_{vdW} + \Delta G_{cav}$) to the solvation free energy is often described with surface area correction, which is proportional to the solvent accessible surface area (SASA), and an additional constant.

$$\Delta G_{vdW} + \Delta G_{cav} = \gamma A + \beta$$

2.3.4 Calculation of relative binding free energies

The ability to study the structural differences on the molecular level and theoretically estimate the effects of structural modification on the binding energy is a powerful method with numerous potential areas of application in current research [205, 206]. Experimental analysis of relative binding free energies has generated thermodynamical data of protein-DNA binding and specificity [40]. Calculations of binding free energies complement experimental data and provide new insight into the macroscopic properties of biochemistry and thermodynamics [215]. Furthermore, the relative contributions of the different components of ΔG can be studied in detail [120]. Theoretical calculations of the relative binding free energies serve as a cornerstone in disease research and rational drug design, however, accurate scoring functions remain a challenge.

Calculation of *absolute* binding free energies (ΔG) is a difficult task, mainly due to the relatively large conformational changes that usually occur upon complex formation [205, 206]. In contrast, calculating *relative* binding free energies ($\Delta\Delta G$) for structurally similar complexes is feasible. Performing such computational and non-physical transformations, usually referred to as *computational alchemy* [102], is a commonly adopted technique for studying what effects structural modifications have on the binding affinity. Thermodynamic cycles and thermodynamic integration are the basis for these calculations and are described in more detail here.

Thermodynamic cycle

The thermodynamic cycle approach can be used for calculating the theoretical values of $\Delta\Delta G$s. A thermodynamic cycle defines the relationship between experimentally measured and computationally calculated free energies (ΔGs). The thermodynamic cycle depicted in Fig. 2.13 is useful for calculating relative binding free energies for a protein binding to two different DNA sequences (one original and one modified sequence, DNA and DNA^M, respectively) in solvent (aq).

As the free energy (ΔG) is a state function, the sum of the energies traversing the cycle has to be zero (disappear). Thus, the $\Delta\Delta G$ can be expressed as:

$$\Delta\Delta G = \Delta G_2 - \Delta G_1 = \Delta G'' - \Delta G'$$

Since only the final states are considered, the transformation itself may be non-physical ("alchemical"). The quantities ΔG_1 and ΔG_2 are measurable in the laboratory, whereas $\Delta G'$

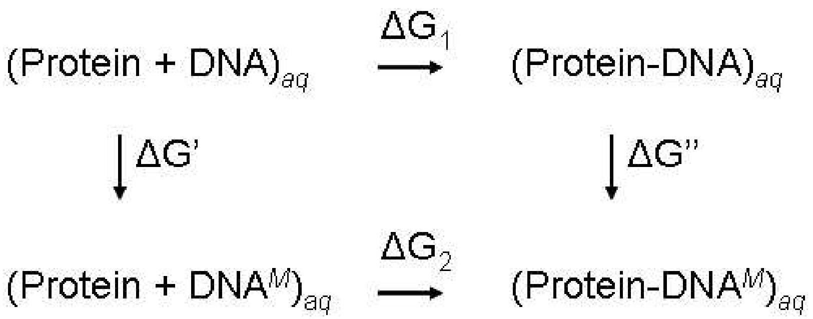


Figure 2.13: The illustrated thermodynamic cycle describes the relationship between free energies that are measurable in the laboratory (ΔG_1 and ΔG_2) and computationally feasible transformations ($\Delta G^{'}$ and $\Delta G^{''}$). Simulating the binding process (represented by ΔG_1 and ΔG_2), compared to the small structural transformation from $protein-DNA$ into $protein-DNA^M$ (represented by $\Delta G^{'}$ and $\Delta G^{''}$), is by far much more computationally intensive.

and $\Delta G''$ are not. In contrast, the non-physical pathways $\Delta G'$ and $\Delta G''$ involve less structural reorganization of the system and are likely to be more reliable from a computational point of view and can be calculated theoretically [148].

Thermodynamic integration method

The thermodynamic integration (TI) method is commonly adopted by computational chemists and molecular modelers in order to calculate the difference in free energy caused by a modification of the system (e.g. the $\Delta G'$ or $\Delta G''$ transformations).

The starting system for the TI method is a hybrid system describing both the starting (*unperturbed*, $\lambda = 0$) and the ending system (*perturbed*, $\lambda = 1$). The mixing parameter λ is used for describing the step-wise transformation going from $\lambda = 0$ to $\lambda = 1$, which is illustrated in Fig. 2.14. The relative free energy ($\Delta\Delta G$) between two systems resulting from a certain transformation is obtained by comparing the ΔGs.

The ΔG between the two states is defined by the following integral:

$$\Delta G \equiv G(\lambda{=}1) \text{ - } G(\lambda{=}0) = \int\limits_0^1 \langle \delta \mathrm{V}/\delta\lambda \rangle_\lambda \mathrm{d}\lambda$$

The integral can be numerically estimated by calculating the sum of ensemble averages of the change in potential energy V ($\langle \delta V/\delta\lambda \rangle$) between consecutive windows i and $i+1$ of λ, using the following equation

$$\Delta G \approx \sum_{i=1}^{n} \delta G_i(\lambda_i \rightarrow \lambda_{i+1}) = \sum_{i=1}^{n} w_i \langle \delta V/\delta\lambda \rangle_{\lambda_i}$$

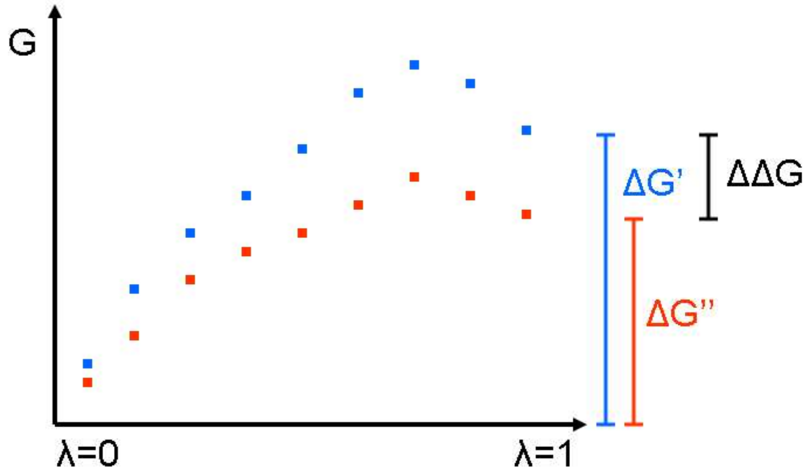


Figure 2.14: The relative free energy ($\Delta\Delta G$) between two systems is obtained by comparing the two individual ΔGs, which are obtained during the transformation going from $\lambda = 0$ to $\lambda = 1$.

In practice, this estimation is achieved by performing an equilibration (for adjusting the structure at the new λ_i) followed by a simulation (for sampling the ensemble average $\delta V/\delta\lambda$) at discrete value of λ going from 0 to 1. The $\langle\delta V/\delta\lambda\rangle$ between two intermediate steps of λ is estimated (sampled) using MD. At each quadrature point (λ_i) the $\langle\delta V/\delta\lambda\rangle$ value is multiplied by a weight w_i that is symmetrical around $\lambda = 0.5$. Several methods exist for estimating the integral e.g. the trapezoidal rule [2, 44].

Ipsa Scientia Potestas Est

Sir Francis Bacon

3 Related work

3.1 Experimental techniques

Rapid technological development, both in terms of efficiency and resolution, has lead to an improvement of experimental methods for analyzing protein-DNA interactions. These methods range from being time-consuming small-scale experiments, such as X-ray crystallography, providing atomic details of the interactions in a specific complex [71], to large-scale experiments analyzing the expression profiles of thousands of genes simultaneously [112]. Theoretically, the larger the collection of binding site data for a certain TF, the more statistically representative the TFBS should be. In practice, however, the experiments are typically performed *in vitro* and experimental artifacts (such as experimental design or not enough sampling of binding sites) introduce noise or bias into the TFBS models [149, 244]. Optimizing the compromise between the coverage and the qualitative aspects of each of the methods remains a key challenge.

Numerous studies with the primary aim to analyze protein-DNA interactions and specific binding preferences have been undertaken. Experimental approaches generate an ever growing body of data available for analysis. The data used in this work comes from a variety of such experiments and provides the basis for all computational approaches presented in Chapters 4 to 6. Here follows a brief description of the most commonly used experimental methods for obtaining both large-scale expression data related to gene regulation and binding specificities of individual TFs.

In situ labeling using microarrays is a technique for studying the expression of thousands of genes simultaneously [160, 220]. Labeled single-stranded probes hybridize with complementary mRNA sequences, which enables the quantitative detection of the transcriptome (the transcribed genes) in a sample. Genes with a similar expression pattern in response to an external signal are often assumed to be biologically related, e.g. for example controlled by a similar regulatory mechanism, and can be extracted and further analyzed [41, 90].

Reporter constructs are employed for analyzing the functionality of whole regulatory regions. Systematic deletion of DNA fragments in connection to a reporter system, will indicate how essential a certain fragment is for the correct expression of a certain gene and how the expression is affected by modifications made to the regulatory regions.

Band-shift assays and gel retardation studies [86, 84] are often used in combination with methods for identifying the TF that binds to a certain DNA sequence, e.g. chromatin immunoprecipitation (ChIP) [117, 209]. Thereby, a quantitative measure of the binding affinity can be obtained.

SELEX and Phage display are two selection experiments for finding a set of preferentially bound DNA sequences [48, 49]. These methods are especially useful for analyzing the level of conservation at each position of the TFBS, since a statistical profile can be derived from the obtained binding sitess [213].

Titration experiments Titration is a standard chemical laboratory method for determining the concentration [mol/L] of a known reactant. A thermodynamical measure of the binding affinity can be obtained by doing titration experiments for a certain TF. The binding strength of the complex is defined by the dissociation constant (K_d). In 2001, Bulyk *et al.* [40] presented a new technology, in which microarrays are employed in order to obtain such K_d:s for several TFs binding to different DNA sequences.

Structural methods: X-ray and NMR Two methods enabling studies of protein-DNA interactions at atomic detail are *X-ray crystallography* [210] and *nuclear magnetic resonance (NMR)* [73]. The two methods differ in a number of aspects. X-ray crystallography provides higher resolution, requires crystals, and is possible for large molecules, whereas NMR can be used for studying molecules in solution, provides information of flexibility, but is restricted to smaller molecules (30 kDa).

3.2 Computational approaches

A vast amount of experimental gene expression data, genomic sequences, and structural data collected in databases [11, 26, 200], has provided the fuel required for computational analysis of several different types of data sets. Theoretical and computational methods provide means for extracting valuable information from such experimental data and contribute to the formalization of hypotheses. Connecting experiments with computational models, and thereafter refining these models through tight feedback coming from further experimental verification is essential for improving the models.

A diverse set of computational milestones, important in all aspects of specific protein-DNA interactions, are presented in the following sections. The two main topics are sequence-based and structure-based methods, which reflect the categorization of the related work presented over the years. Common to all methods is to represent specific interaction preferences between proteins and DNA, thereby modeling the interactions and binding affinity. This is important in order to enable detection and discovery of TFBSs on a genomic scale and to understand the effects mutations have on the binding affinity in protein-DNA complexes. The TFBS models serve as a key component for elucidating potential transcriptional mechanisms, function, and for modeling regulatory networks. The potential arising from the ability to reliably predict functional and high-affinity binding sites is enormous and a trend towards integrating sequence-based and structure-based information has been observed [105].

3.2.1 Sequence-based methods

The immense amount of sequence data available for several organisms has led to a rapid development of sequence-based analysis tools. The genomic DNA sequences are interesting from a gene regulatory perspective, since these contain the information needed to direct transcription. The TFBSs are short sequences, typically located in the non-coding DNA close to the affected gene. Predicting genes for various organisms is clearly a non-trivial task [194]. Biological information, e.g. the presence of signals around the TSS, bp statistics (coding vs. non-coding regions), C_pG islands, and the presence of TFBSs [27, 98, 274],, for predicting promoters. A detailed analysis of the promoter structure provide clues about potential regulatory networks and common biological functions for groups of genes [201].

Representation and discovery of TFBSs are closely related challenges [244]. Reliable models for detecting TFBSs are needed e.g. for promoter prediction. The most basic representation of a TFBS is the *consensus sequence*, which is a sequence stating the most preferred nucleotide at each position of the binding site for a certain TF [244], see Table 3.1 for an example. Such a consensus sequence, is usually obtained by aligning a collection of known binding sites, thereby constructing a representative consensus sequence, which presents the most likely nucleotide at each position. The consensus sequence (used as a regular expression) was the

Table 3.1: Two sequence-based representations of the same TFBS. The consensus sequence and the position specific scoring matrix (PSSM) representations of the Zif268 TF.

Position		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Consensus		T	G	C	G	T	G	G	G	C	G
PSSM	**A**	5	7	0	2	0	31	0	0	13	0
	C	3	0	*98*	0	2	0	0	0	*76*	0
	G	5	*93*	0	*98*	14	*69*	*100*	*100*	0	*100*
	T	*87*	0	2	0	*84*	0	0	0	11	0

first TFBS representation to be used for scanning sequences on a genomic scale [34, 80, 245, 260]. However, a consensus sequence is not a satisfying representation of a TFBS, since it usually allows for variations in some of the positions. The *position specific scoring matrix (PSSM)* is a refinement of the consensus sequence, since it captures the statistical occurrence (frequency) of each nucleotide at each position of the binding site. An example of a PSSM is illustrated in Table 3.1. PSSMs are derived in a similar manner as consensus sequences, i.e. by collecting and aligning known binding sites, and can be represented as a sequence logo shown in Fig. 3.1 [282]. In a sequence logo the total height of the bases at each position is the information content (IC) of the position, and the height of each individual base is the proportion of that base of the total height. The IC of a position i in a TFBS is defined as:

$$IC(i) = \sum_{b=A}^{T} P(b,i) \times ln \frac{P(b,i)}{P_{ref}(b)}$$

where $P(b,i)$ is the probability to observe base b at position i in the binding site. The IC is the average specific binding energy, as defined above, and will be zero if the bound probabilities are the same as the prior probabilities (i.e. a non-specific position has zero IC).

Consensus sequences and PSSMs can be used for efficient scanning of genomic DNA sequences for potential binding sites. However, these models are very sensitive and suffer from severe problems in specificity, often predicting a high number of false positive TFBSs [244]. The reliability of the models are in many cases limited by the number of experimentally verified sites and the lack of a quantitative measure of the sites that have been shown to be functionally active.

Sites contributing to a PSSM can also be derived in a purely computational fashion. This approach is based on the observation that TFBSs are enriched in regulatory sequences [76, 158].

The regulatory regions are extracted and used as input for algorithms searching for sta-

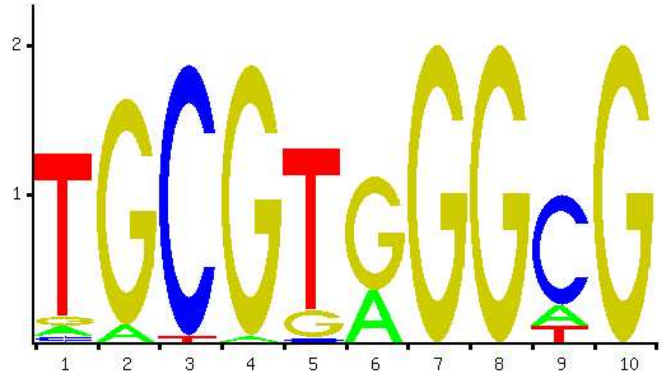


Figure 3.1: The PSSM can be converted into a sequence logo using the information content (IC) at each position. A well conserved position is represented with 2 bits of information, whereas zero IC means that no base is preferred over the others at that position. The sequence logo of Zif268 (the consensus sequence and PSSM are showed in Table 3.1) is illustrated in this picture.

tistically over-represented motifs of a given length [244, 252, 275]. Gibbs sampling [146], ANN-Spec [281], and Multiple Expectation maximization for Motif Elicitation (MEME) [10] are good examples of motif discovery tools. An over-represented motif is represented as a PSSM and is considered to be a fairly good representation of the true binding site [244].

The identification of genomic sequences, which have been conserved throughout evolution (comparative genomics), received considerable attention as a means for refining the search space for the sampling algorithms. Genomic sequences from two or more species are used for cross-species comparisons to find conserved regions, which are more likely to be functional as TFBS than other less conserved regions. This approach - *phylogenetic footprinting* - restricts the search to conserved regions and has improved the motif discovery process immensely and eventually led to a more efficient way for delineating potential regulatory mechanisms [28, 68, 268]. The evolutionary distance between the compared species is important. *H. sapiens* and *M. musculus* are about 60 Myrs apart, which appears to be enough time for the genomes to show differences in the non-coding sequences and share reasonable similarity of functional sequence elements [55, 288].

Additional biological knowledge has been integrated into the prediction methods in order to enhance the usefulness and reliability of the predictions. The simultaneous detection of binding sites of functionally related TFs, helps in the identification of so called regulatory modules [142, 267, 274]. Selecting a set of target genes, for which a common regulatory mechanism is to be expected, is a well-explored possibility to improve the discovery of individual functional TFBSs [197, 218, 267] and combinatorial regulatory modules consisting of several cooperating TFBSs [97, 128]. Obtaining functional clusters of genes using microarrays is a commonly applied technique and shows good results [41, 110]. This approach of combining functional with regulatory analysis is useful in many aspects [54, 100, 112].

The methods discussed above all rely on the PSSM models for detecting the TFBSs and regulatory modules. It has been shown that the most basic assumption used by the sequence-based methods, that the nucleotides contribute additively to the binding energy, is incorrect [22, 39]. The assumption about additivity is, however, a good approximation for some TFs. Attempts to model dependencies between neighboring bases, by including higher-order effects into the models have been reported [14, 133, 284] and show an improvement over additive models. The binding of one TF domain (e.g. in multi-ZF proteins) is dependent on the binding of the other neighboring domains [47, 113]. Furthermore, synergy effects also occur between TFs that do not bind to each other when binding to a regulatory DNA sequence [97, 99].

The representation and discovery of TFBSs, the analysis of regulatory regions, the construction of potential regulatory modules, and the analysis of functionally related genes has undergone a remarkable improvement and has provided means for studying regulatory networks in a wide range of organisms [112, 197, 251]. Sequence-based methods have a few clear limitations. Side Chain flexibility at the binding interface, differences in molecular conformation and stability, and hydration at the binding interface are hard to model, thus typically not included in sequence-based approaches.

3.2.2 Structure-based methods

The sequence-based models used for representing the TFBSs are still being questioned for their correctness and applicability. The whish to define a general, or at least a family specific, recognition code explaining specific protein-DNA interactions (discussed in Sect. 2.2), further pushed for the integration of additional experimental data and biological knowledge into the TFBS models [105]. The information residing in structural data, immediately appeared to be the missing link when going from experimental observations to the formulation of hypotheses regarding the interactions [126]. The contacts between the TF side chains and the nucleotides in the binding site can be studied at atomic resolution by analyzing protein-DNA co-crystal structures [71, 166, 130]. Such structural data provide insights into the complex network of intermolecular interactions, however, provides no information about the dynamics of the contacts.

Threading approaches

Structure-based methods aim to establish empirical contact potentials between protein and DNA from structural data [141, 168]. These potentials are similar to potentials used for threading and evaluating protein structures [171]. These knowledge-based interaction potentials are used to describe the interaction preferences and can be used for efficient scanning of genomic DNA sequences. The underlying contact probabilities are collected from a data set containing structures of TF-DNA complexes representing different structural families of TFs [165]. Pair-wise contact tables are generated by regarding contacts within a certain radius as described in Fig. 3.2. By studying mutant structures, it is possible to retrieve information about how flexible some of the interacting residues and bases are [177]. An approach to structurally align the protein-DNA interfaces in order to uncover spatial relationships and comparing the geometry of the interfaces for better prediction was presented recently [233]. One limitation is that some TF families are not represented, due to difficulties associated with experimental protein purification. The structure-based empirical potentials in combination with defined contact tables can be used for predicting TFBSs on a genome-wide scale and have shown to be superior to sequence-based PSSMs [141, 168]. The additivity approximation is often adopted, but dependencies between the interactions can be modeled to some extent [23]. The main advantage of the structure-based approaches is that amino acid-base propensities and structural orientations are included in the models [195, 285]. The empirical potentials can be established for subsets of the data (e.g. TF families) or for defining different types of interactions, such as specific (amino acid-base) and non-specific (amino acid-backbone). A disadvantage is that crucial water molecules at the interface are typically neglected by these potentials, as they are usually excluded in the extraction of the contact tables and construction of the empirical potentials.

Molecular modeling approaches

Molecular modeling holds great promise for addressing the main three issues where the sequence- and non-dynamic structure-based approaches fail, namely structural re-arrangements (flexibility), interdependence (non-additivity) [285], and hydration effects at the protein-DNA interface [58, 147]. Molecular dynamics simulations allow for analysis of the flexibility and thermodynamics of a protein-DNA complex over a short period of time, rather than the rigid snapshots that structural data provides [140]. These flexible models can theoretically

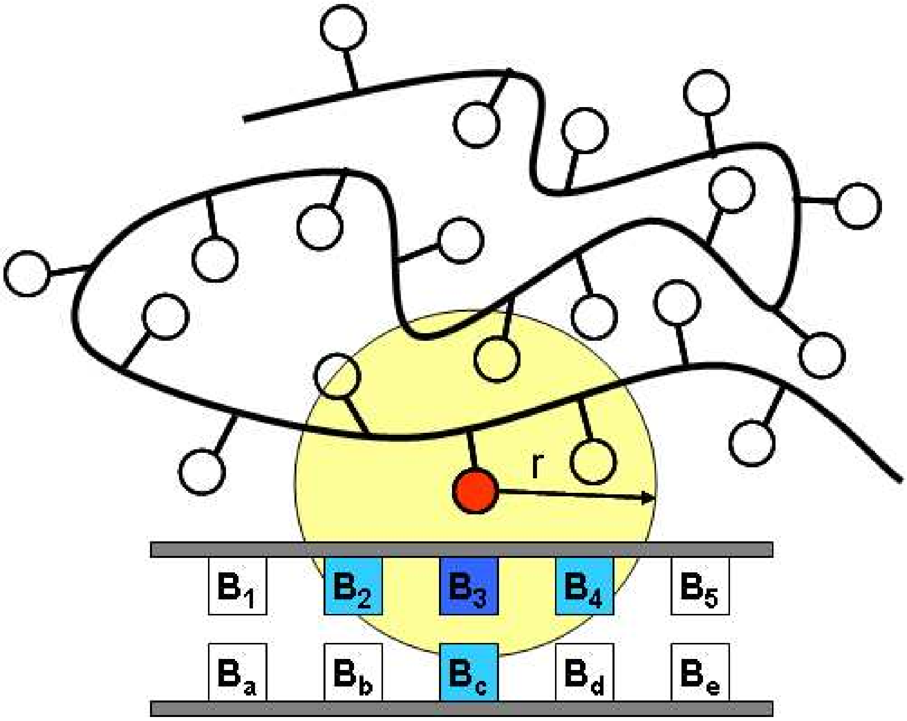


Figure 3.2: Statistical potentials are obtained by generating pair-wise amino acid base contact tables from structural data. All contacts between an amino acid side chain (red) and bases (blue) within a maximum radius (r) are used for generating the statistical potential.

be considered a more truthful representation of the biophysical reality of the interactions, however, suffer severely from high computational requirements. Approximations imply a decreased accuracy, but are possible for several aspects. For biomolecular simulations such approximations contribute to the necessary speed-up of the computations, thereby enabling large-scale evaluation of several complexes. Restrictions in the flexibility by applying force constraints during the simulations [72], trying a set of specified conformations by means of side chain placement and structural alignment [233], or restricting the search space by Monte Carlo/Umbrella sampling [148, 221] are commonly employed techniques.

Simulations of fully flexible systems in explicit solvent are time-consuming but possible for investigating individual complexes [92, 127]. The effects that small chemical modifications, made to functional groups of the nucleotides or the amino acid side chains, have on the binding affinity can be studied in this way [134, 215, 228]. Observing the resulting conformational changes or the changes in relative contributions to the binding free energy provide useful insights into the physical reality of the direct and indirect interactions [120]. Furthermore, it allows for a detailed analysis of the hydration by monitoring essential water molecules trapped at the interface [256]. A further commonly adopted technique is to use an implicit solvent model [139] for estimating the solvation free energies (discussed in Sect. 2.3.3). Implicit solvent models are suitable for analyzing the effects of larger modifications, such as whole amino acid side chain mutations or bp mutations. As computations using approximate models are faster than explicit models they can be used for analyzing a larger set of complexes.

Molecular dynamics is a time-consuming and computer intensive way to improve the current understanding of protein-DNA binding. Nevertheless, it seems to provide some missing clues about the thermodynamics and flexibility that are impossible to retrieve form other data [72, 105]. Choosing the appropriate approximations is highly dependent on the type of study and should be considered a problem-specific decision.

To acquire knowledge, one must study; but to acquire wisdom, one must observe.

Marilyn vos Savant

4 Gene regulation at the systems level

Each year 10.9 million people are diagnosed with cancer and 6.7 million people die from various types of cancer worldwide [1]. Normal regulation of genes permits the development and differentiation of healthy organisms, whereas abnormal gene regulation can lead to serious diseases such as cancer.

Tumor suppressor genes (proto-oncogenes) are genes normally present in cells, where they exert essential physiological functions. Genetic alterations, such as mutation, translocation, or amplification, can turn these into oncogenes. For example, oncogenes can possess the ability initiate cell division and disturb the normal cell cycle control mechanisms, which can lead to an uncontrolled cell proliferation. Several key TFs have been identified as oncogenes, thereby playing an active role in tumor progression. Cancer is considered to be a gene regulatory disease, since cancerous cells are controlled by abnormal gene regulatory mechanisms.

Analyzing the causes underlying cancer genesis for different cancer types, can be performed on a large-scale basis at the level of systems. A wide range of experimental techniques, targeting different aspects of tumor development, are used for this purpose. Not only genetic, but also environmental factors, such as radiation and toxins, are known to be involved in cancer development. In general, any one defect is not enough to initiate cancer on its own, the defects that cause cancerous cell growth are typically accumulated over time.

The enormous impact of cancer has governed an intensive research interest and the multifactorial nature of cancer has led to a rapid increase in the volume of experimental data related to cancer. It is important to view the differences between normal and pathological gene regulation at the level of systems, in order to understand the gene regulatory networks that control gene expression.

Analysis of heterogeneous cancer-related data presents a major challenge to computational biologists. Here, a novel approach for analyzing cancer-related data at the level of systems is described, which has proven useful for shedding light on the cancer genesis process. It is demonstrated that differential gene expression correlates with antibody response in several human cancer types and provides evidence for the involvement of gene regulation in cancer development. Thereby, putative target genes involved in the development of several cancer types, including melanoma and breast cancer, are identified. Understanding the different stages of progression from normal to cancerous tissue has important implications for early diagnosis and target gene identification [254].

4.1 Integrative analysis of cancer-related data using CAP

4.1.1 Introduction

Gene regulation precisely controls cell growth and division in living organisms. Normally, replication of the genetic material occurs in a highly organized step-by-step fashion, in which repair enzymes ensure that the daughter cells receive the complete genetic material from the parent cell. Strict control mechanisms monitor the subsequent cell division. Specific DNA-binding proteins, the TFs, play essential roles at each checkpoint of the cell cycle. Cells carrying damaged DNA that cannot be repaired or do not fulfil the criteria at each checkpoint, receive signals to undergo programmed cell death (apoptosis). Cancerous cells show a deviant behavior by being able to escape these apoptosis signals. The normal cell cycle regulation is often disrupted in cancer. A failure to repair damaged DNA or to prevent it from being replicated leads to the accumulation of genetic defects, which eventually can lead to the development of tumors. Most types of cancers result from several co-occurring events, such as genetic alterations, disturbance of signal transduction, or failure of immunological surveillance [176].

The multifaceted aspects of cancer development are reflected in the heterogeneous data that is stored in cancer-related databases throughout the community (a selection is listed in Table 4.1). These databases usually focus on specific fields of cancer research, whereas the complexity of cancer genesis requires an integrated analysis of these data.

Data integration and modeling is a true challenge in bioinformatics and the essential key to being able to answer questions at a systems biology level [46]. The cancer-associated protein system (CAP), which was developed in this work [67], is a novel integrated database and analysis tools for cancer-related data. CAP was designed to enable simultaneous and integrated analysis of data originating from several information sources representing different aspects of cancer development. User-specific experimental data can be imported into CAP, where it is automatically annotated with information from external sources and predictions from integrated bioinformatic tools. The data contained within CAP can be accessed through a web-interface, through which it can be viewed and edited by the user. A further important function is statistical analysis of the user-defined data sets. The results of such analyses can be rendered as graphs and exported as tables for further use.

The correlation between the immune response, genetic alterations, and gene regulation in cancer has been hypothesized. CAP facilitates the analysis of such complex relationships and

Table 4.1: Data sources integrated into CAP and their URLs.

Data source	**URL**
Cancer GeneticsWeb (CGW)	http://www.cancerindex.org/geneweb/
CIDB	http://www2.licr.org/CancerImmunomeDB/
LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
NCBI	http://www.ncbi.nlm.nih.gov/
NCI60	http://genome-www.stanford.edu/nci60/
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/
SEREX	http://www.licr.org/SEREX.html
SWISS-PROT	http://www.ebi.ac.uk/swissprot/

was used for analyzing two different cancer-related data sets. The obtained results indicate that there is a correlation between immunogenicity and over-expression.

Here follows a description of the overall design of CAP and a brief description of its abilities. The results of the analysis will be presented in more detail. CAP is publicly accessible at http://www.bioinf.uni-sb.de/CAP/ and described in detail in [67].

4.1.2 Materials and methods

The main computational challenge in developing CAP is the integration of heterogeneous data, in order to facilitate an analysis process that covers several aspects of cancer development. The content of CAP originates from several sources, the main contributing ones and their URLs are listed in Table 4.1. There are three main types of information sources: experimental data, external sequence information, annotations, and predictions. How these have been combined is illustrated in Fig. 4.1.

Data sources

Two major sources of genomic and proteomic information are the NCBI's Reference Sequence Project (RefSeq) [200] and the SWISS-PROT database [11]. RefSeq provides non-redundant cDNA, genomic DNA, and protein sequences, whereas the focus of SWISS-PROT is protein sequences. Both sources provide a rich body of sequence specific annotations including mutations, polymorphisms, structural domains, post-translational modifications, function, and subcellular localization.

Tumor specific antibodies can be detected using the experimental method SEREX (serological analysis of recombinant cDNA expression libraries [257]). SEREX-related information

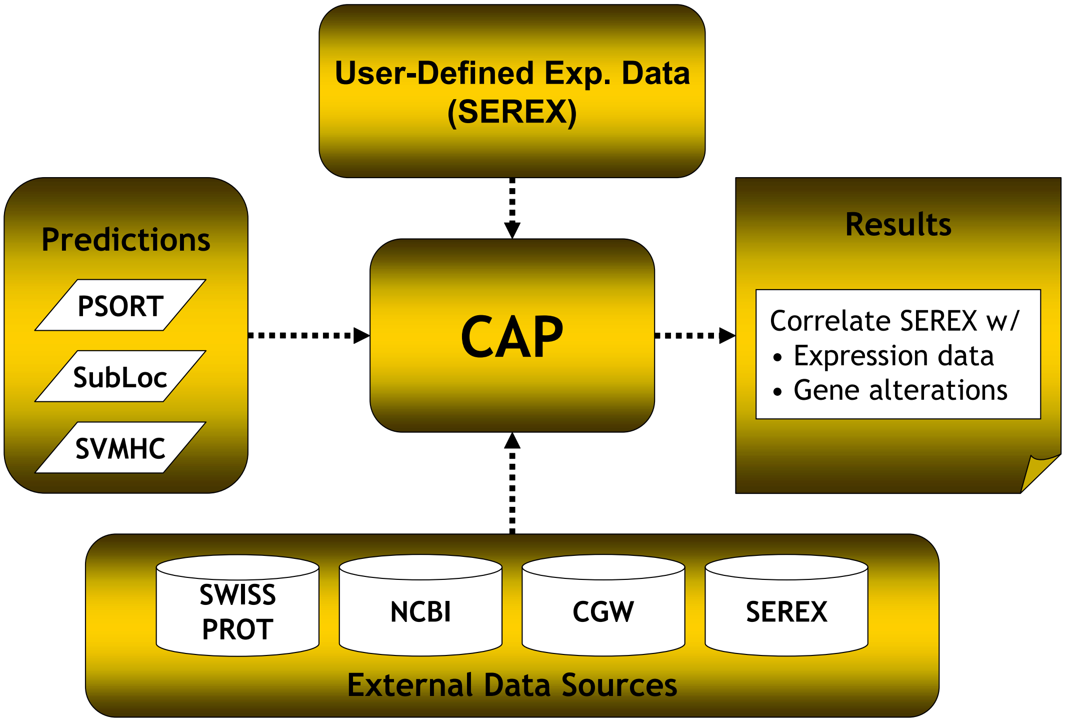


Figure 4.1: An overview of how different data sources have been integrated in CAP in order to facilitate analysis of heterogeneous data.

in CAP comes from own experiments [52] and the SEREX database, which is now contained within the Cancer Immunome Database (CIDB, Table 4.1).

CGW provides information about genetic abnormalities, e.g. mutations, which is related to different genes and listed according to tumor type. The information used in these studies was obtained from CGW and the status of CAP as of December 1, 2003. External database information is updated on a regular basis.

Predictions

Bioinformatics provides a fast mean to computational analysis of both gene and protein sequences. A number of different tools, including prediction of protein subcellular localization, function, and MHC class I epitopes, have been integrated into CAP. The information retrieved from external resources is by no means complete; hence these prediction methods are useful for assigning missing features with a fairly high reliability.

The subcellular localization and function of a protein is important information for understanding potential interaction partners, the participation in metabolic pathways, and role in cell-signaling cascades. The two prediction methods integrated into CAP, are PSORT [182] and SubLoc [108]. The accuracies of the predictions have improved significantly during the

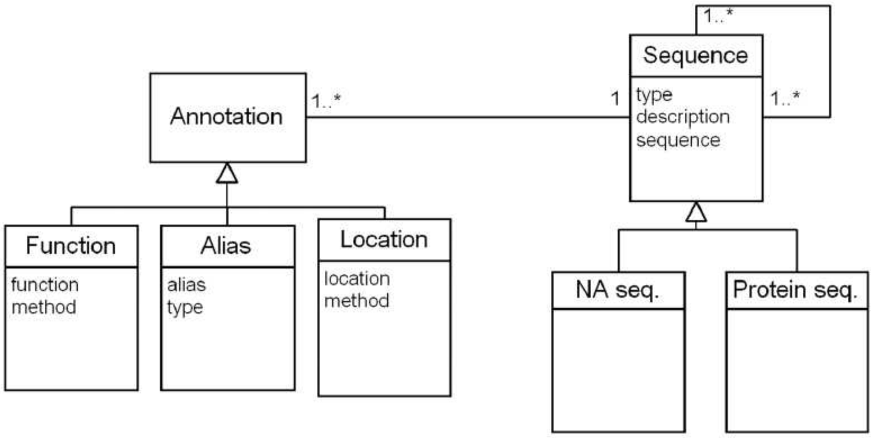


Figure 4.2: An illustration of a part of the CAP data model. The sequence class is a generalization of protein and NA sequences. Annotations of different types can be linked to the sequences. Sequences can be linked to other sequences (e.g. protein to gene), which enables flexible queries across the CAP data model.

past few years [66], the most recent advances are not integrated in CAP to date but can be used externally [104]. Predicting protein function is not a trivial task, regarding the wide spectrum of possible functions, the existence of multitask proteins, and the various ways of defining function. Protein function is predicted by ProtFun [121] and the annotations are added to the CAP database. The presence of putative T-cell epitopes in a protein serves as good candidates in the development of cancer therapies. The prediction of such epitopes is performed using SVMHC [65], which has been integrated into CAP.

Data model and data inspection

The CAP data model was designed to facilitate fast and flexible analysis of different types of heterogeneous biological data. The model is described using the unified modeling language (UML) [214]. A part of the CAP data model is shown in Fig. 4.2. Protein and gene sequences are at the center of the model and can be of two types, either experimental (e.g. from a SEREX experiment) or reference (e.g. SWISS-PROT). The linking to external databases, shown in Fig. 4.3, facilitates fast import and frequent up dates of the content in CAP. An important feature of the data model is that each sequence can be assigned different types of annotations, such as experimental information or results from prediction methods. The key concept in CAP is to integrate heterogeneous cancer-related data and methods to keep track of links between different types of sequences and the annotations belonging to these.

The homogeneous data model allows for a fast and flexible analysis and generic analy-

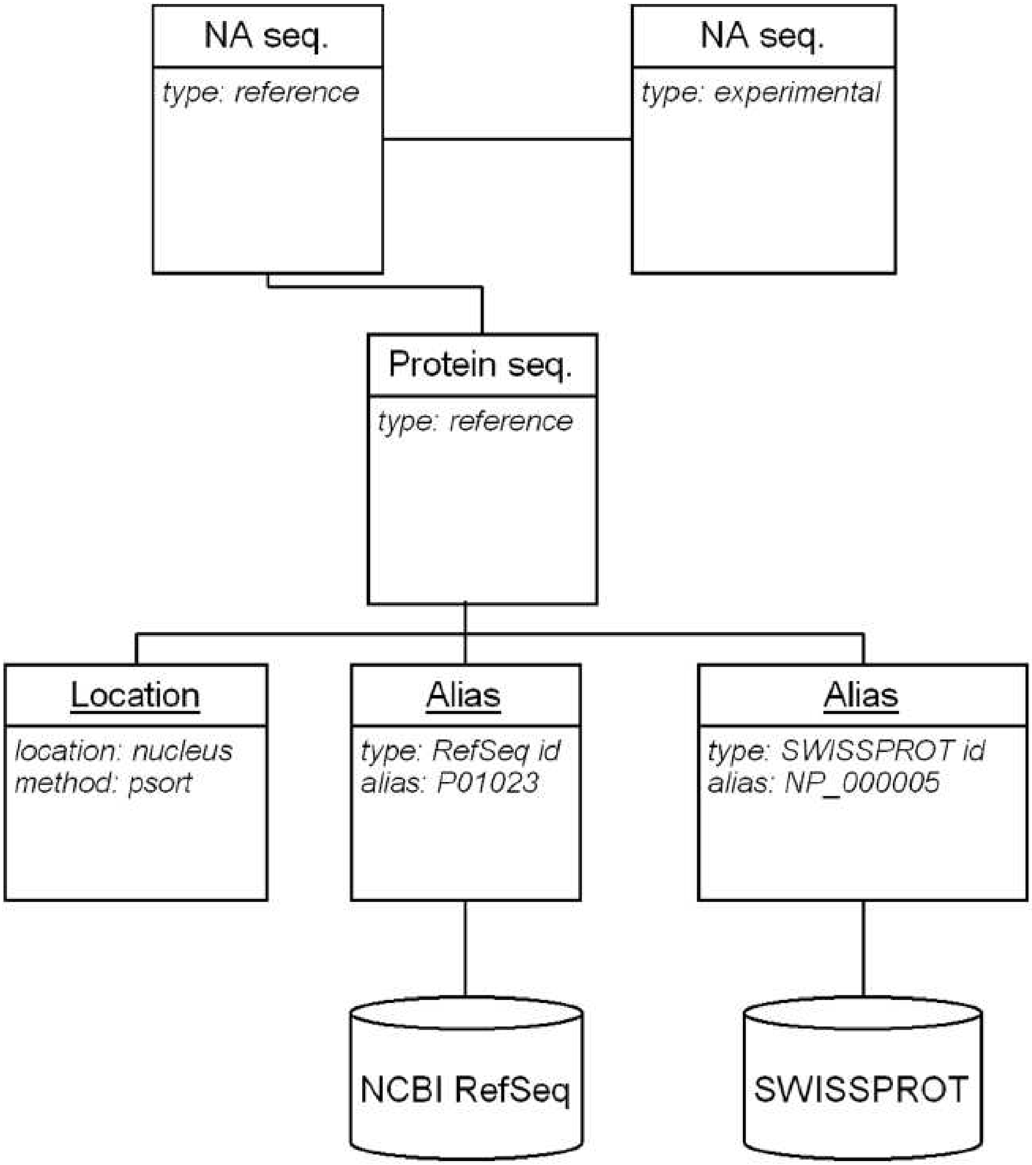


Figure 4.3: External information is extracted and used for annotating sequences in the CAP database. CAP is linked to external databases like RefSeq and SWISS-PROT, which provide valuable sequence information.

sis tools for statistical analysis of the database content. New data can be imported and the data can be grouped into experiment specific data sets. Furthermore, CAP provides a palette of options for analyzing individual sequences or sets of sequences in detail. Statistical information, including chromosomal distribution of genes, protein function, and subcellular localization, can be generated, rendered, and exported for any data set of interest.

4.1.3 Results

The applicability of CAP is exemplified by two large-scale analysis processes. As a first example, the connection between genetic alterations and immunogenicity is analyzed. The second example illustrates how the correlation between immunological data and gene expression data can be analyzed using CAP. These two examples highlight the usefulness of CAP by integrating several important aspects of cancer. The second indicates a correlation between disturbance of the gene regulatory mechanisms and immunogenicity in cancer.

In order to analyze the relationship between gene products carrying known mutations and the development of an autoimmune response in different cancer types, data was collected from SEREX [257] experiments and Cancer GeneticsWeb (CGW), see Table 4.1 for links. A total of 723 genes in CAP have been found by SEREX experiments, 17 genes and two splice variants out of these overlap with CGW. Regarding only those where the cancer types agree in both SEREX and CGW, seven genes were extracted. The genes TP53 and GSTT1 (glutathione S-transferase theta 1), were among these seven, and are known to carry specific mutations or polymorphisms. Interestingly, TP53 has previously been found to cause an immune response and carry mutations in both colon and breast carcinoma [6, 234]. Mutations in TP53 have been found is several further cancer types, however, no immune response has been shown. As for GSTT1, antibody responses occur in patients with breast cancer. This tumor is associated with specific GSTT1 polymorphisms [179]. However, these types of polymorphisms also occur in other tumors including head and neck cancer without an antibody response [45]. Other examples of genes include NME2/NME1 (protein NM23B/A expressed in non-metastatic cells 2/1), HSPCA (heat shock 90kD protein 1), Ki-67 (MKI67), and MIF (macrophage migration inhibitory factor). NME1 and NME2 have been reported as immunogenic and over-expressed in malignant colon carcinoma [172], HSPCA in renal cell carcinoma [185], Ki-67 in melanoma [103, 101], and MIF in melanoma [231].

In the second application of CAP for analyzing cancer-related data, the correlation between over-expression and an autoimmune response was analyzed. The gene expression data of the

NCI60 microarray project was used for extracting over-expressed genes. The NCI60 data set (link in Table 4.1) contains the expression profiles of 8,000 genes from 60 cancer cell lines. These cell lines are also used by the National Cancer Institute (NCI) for screening potential cancer drugs. Two criteria have to be met for genes to be considered over-expressed, the genes have to show at least a two-fold expression level and have measured expression levels in at least four of the 60 cell lines. In total, 319 genes occur in both CAP and the NCI60 data set; 277 of these genes were over-expressed in at least one of the NCI60 cell lines, while 69 were over-expressed in at least 10% of the cell lines. A total of 13 genes are found to be immunogenic in SEREX experiments, are over-expressed in at least three tumor specific cell lines, and are of the same cancer type. These 13 genes and SEREX related information are presented in Table 4.2. There are five genes related to melanoma (COL9A3, HEXB, RRBP1, SLC2A11, and TIMP3) that are immunogenic and over-expressed. Furthermore, the list contains genes from other common cancer types such as breast, colon, lung, and renal cancer.

4.1.4 Discussion

The immune response plays an important role in disease development, since it is the body's own defense mechanisms. One of the new approaches to the treatment of cancer is to up-regulate the activity of the immune system in order to allow it to better control carcinogenesis (also known as immunotherapy [3]). It has been hypothesized that the antigens causing the immune response stem from genes that are altered by tumor-specific mutations or have a changed expression profile in a certain tumor (reviewed in [176]).

The analysis of the current data suggests that immunogenic antigens in cancers are likely to occur as a result of over-expression. There is no clear evidence to be seen, that an immune response is triggered due to genetic alterations, such as mutations or polymorphisms. The analysis presented in this work is by no means complete and further experimental analysis is needed, therefore these results should not be interpreted as rules regarding cancer development. Furthermore, genetic aberrations in certain chromosomal regions and changes in gene-expression of such regions have been identified for certain cancer types [18]. Over-expressed genes have also been suggested as attractive therapeutic targets [289].

The two applied examples illustrate that CAP facilitates this kind of integrative analysis of large-scale cancer-related data originating from the fields of genetics, proteomics, and immunology. As new experimental techniques are developed and the body of data grows, it

Table 4.2: SEREX-related information for the 13 genes found in the same cancer type in both SEREX experiments and over-expressed in the NCI60 data. For a number of genes several related SEREX clones were found.

Abbreviation	Gene name	RefSeq id	SEREX clone	SEREX tumor
Melanoma				
COL9A3	collagen, type IX, alpha 3	NM_001853	Hom.TsMe3-89	melanoma
HEXB	hexosaminidase B (beta polypeptide)	NM_000521	Hom.TsMe2-12	melanoma
RRBP1	ribosome-binding protein 1 homolog 180kDa	NM_004587	TE53	unclassifiable
			TM-76	melanoma
SLC2A11	solute carrier family 2 (facilitated glucose transporter), member 11	NM_030807	Mz19-64	melanoma
			NY-SAR-47	fibrosarcoma
TIMP3	tissue inhibitor of metalloproteinase 3	NM_000362	Mz19-3	melanoma
Breast cancer				
P8	p8 protein (candidate of metastasis 1)	NM_012385	NY-BR-89	malignant breast
TP53	tumor protein p53 (Li-Fraumeni syndrome)	NM_000546	NY-Co-13	colorectal adenocarcinoma
			NY-BR-94	malignant breast
			NW-F14	malignant colon
			NW-F93	malignant colon
CENPF	centromer protein F, 350/400ka (mitosin)	NM_016343	MOC-SW-139	malignant colon
			MOC-SW-18	malignant colon
			MOC-SW-151	malignant colon
			NGO-Br-7	malignant breast
			MO-TES-148	malignant colon
			NY-ESO-11	esophageal cancer
			NGO-Pr-24	malignant prostate
			NY-BR-69	malignant breast
GBP1	guanylate binding protein 1, interferon-inducible, 67 kDa	NM_002053	NGO-Br-40	malignant breast
Colon cancer				
SCNN1A	sodium channel, non-voltage-gated 1 alpha	NM_001038	NW-CD35b	adenocarcinoma colon
AP1G2	adaptor-related protein, complex 1, gamma 2 subunit	NM_003917	NW-SW15	adenocarcinoma colon
Lung cancer				
TRAP1	heat shock protein 75	NM_016292	LC19	malignant lung
Renal cancer				
PFKFB3	6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 3	NM_004566	NY-REN-56	malignant kidney

becomes more and more important to design databases that offer more than simple data retrieval [38]. Tools for predicting and performing statistical analysis of subcellular localization, function, and T cell epitopes have been implemented into CAP. The automatic annotation of the information contained in the database is an illustrative example and serves as a base for exploring protein interaction networks and signaling pathways involved in cancer.

Integration, annotation, and large-scale analysis of heterogeneous data help in the understanding of genetic and immunological basis of cancer genesis. The novel analysis pipeline CAP is one of the first of its kind, offering large-scale analysis tools for heterogenous data sets. The results include evidence for the involvement of gene regulation in cancer development. A global analysis of this type is necessary in order to be able to identify sets of biologically related genes. The specific regulation of such sets is to analyzed by zooming in to the sequence level.

5 Gene regulation at the sequence level

The cellular processes that can be observed at the level of systems are orchestrated by gene regulatory networks. Sequence-based methods provide powerful means for analyzing transcriptional regulation, as these make it possible to zoom in on specific regulatory sequences and the TFBSs therein. The regulation of biologically related genes and the functions of those genes can be studied in detail. In this chapter, important gene regulatory control mechanisms in three different organisms are analyzed using sequence-based methods. The knowledge acquired at the level of sequences is helpful for understanding regulatory effects at the level of systems. Furthermore, sequence-based analysis of gene regulatory regions is a crucial step towards reconstructing gene regulatory networks.

First, an analysis of the regulatory mechanisms controlling human melanoma-related genes that were identified at the level of systems using CAP (described in Chapt. 4) is presented. Potential TFs involved in melanoma are suggested. Putative binding sites for the TF AP2 are detected, which is further confirmed by experimental evidence showing that AP2 plays a key regulatory role in melanoma development. In the second study, the system for integrative analysis of yeast regulatory sequence analysis system (YRSA) is described. The regulation of biologically linked genes is investigated in three case studies using YRSA. Putative functional TFBSs have been identified in a set of regulons and the TFs likely to bind these are suggested. Especially, the TF MCB1 (or a TF binding to a highly similar TFBS) is suggested to be involved in DNA damage response. Finally, an analysis targeting the regulatory network determining stem cell fate and floral patterning in *A. thaliana* is described. High significance motifs are presented and TFs putatively binding to those are suggested. These analyses show that it is possible to map out TFs responsible for certain cellular traits and to continue with functional characterization of the target genes.

5.1 Regulatory analysis of melanoma-related genes

5.1.1 Introduction

An indication of that over-expression can cause an immune response in cancer was found using CAP [67]. At least five genes in melanoma: COL9A3, HEXB, RRBP1, SLC2A11, and TIMP3, listed in Table 4.2, show this behavior. Over-expression of a gene in a cancer cell compared to a normal cell can be the result of a disturbance of the transcriptional control. Due to the nature of the regulatory networks underlying transcription and the central roles of TFs, it is likely that several genes are affected by such a change simultaneously [267, 218]. The aim of this study is to elucidate putative commonalities in the transcriptional regulation of the five melanoma-related genes.

Computational analysis of the transcriptional regulation of human genes presents many challenges. The human genome is not only large, but the coding and functional information contained within is widely spread out in the non-coding and repetitive sequences (background). The exons of genes are interspersed with long introns and non-coding (functional) regulatory regions often stretch out over several kb long sequences (illustrated in Fig. 2.3). Potential mechanisms underlying the regulation of these five melanoma-related genes and possible causes of the over-expression in cancer are addressed in this sequence-based study. The genomic sequences upstream of the coding cDNA sequences for the five human genes and the corresponding sequences for the mouse (*M. musculus*) genes were extracted. Using phylogenetic footprinting, the upstream sequences of two suitably diverged species (like human and mouse) are aligned in order to highlight conserved and potentially functional regions in the non-coding DNA. The promoters of the genes were extracted and the conserved sequences were searched, using a Gibbs sampling approach [146] and MEME [96], for potentially over-represented motifs. These over-represented motifs are compared against motifs in the JASPAR database [217] containing experimentally verified motifs. Good matches potentially correspond to known functional TFBSs. The detected motifs were also checked for their statistical significance in the regulatory regions. The preliminary results show that there are a few interesting motifs, which share a high similarity to experimentally verified TFBS motifs. Interestingly, the melanoma genes have a few individual motifs and do also share a few common motifs in their upstream regions, which are previously known for their involvement in cancer. Furthermore, the potential involvement of a few new TFs in the development of melanoma is discussed. A detailed analysis of this kind can highlight critical aspects of regulatory mechanisms behind the development and progression of cancer. There is a clear need for further experimental verification of the results. Nevertheless, they are likely to prove useful for understanding the development of this cancer type and in the process of identifying putative drug targets.

5.1.2 Materials and methods

Sequence-based analysis of gene regulatory regions rely on several consecutive steps, for a graphical illustration see Fig. 5.1. The data sources, tools, and their URLs used in this stepwise analysis are listed in Table 5.1 and additional details are provided in Sect. 5.2.2. The genomic and cDNA sequences of the human and their homologous mouse genes were obtained using the NCBI's RefSeq database and the Ensemble Genome Browser. One of the genes, SLC2A11 has no corresponding sequence in mouse and was excluded from further analysis. The genomic sequence was aligned to the cDNA sequence in order to identify the TSS, for human and mouse respectively. After the TSSs have been identified, the genomic human and mouse DNA sequences were aligned using ClustalW [253] to identify conserved sequences in the upstream non-coding regions. Conserved sequence segments (matching requirements on the conservation level of 65% in a sliding window length of 100 bps) in the region between five kb upstream and the TSS were extracted for further analysis. These conserved sequences were fed into a Gibbs sampling algorithm and into MEME, which were the tools used for identifying statistically over-represented sequence motifs. At the final step of the process the resulting motifs have to be analyzed. The motifs found using the two sampling approaches were checked against a collection of reference motifs in the JASPAR database [217], in order to assign potential identities to the anonymous motifs that result from the sampling. The significance of the over-representation was statistically verified by comparing the occurrence in a) the conserved regulatory regions, b) the full five kb upstream regions of the melanoma-related genes, and c) the five kb upstream of all genes in the genome. In total, five motifs were selected for further investigation, which involved manual annotation and literature studies.

5.1.3 Results and discussion

Five human cancer-related genes found to be both immunogenic and over-expressed in melanoma using CAP, four out of these five genes were used in this study: COL9A3, HEXB, RRBP1, and TIMP3. The transcriptional regulation of these genes was analyzed using common sequence-based methods such as: sequence alignment, extraction of conserved regions, sampling of

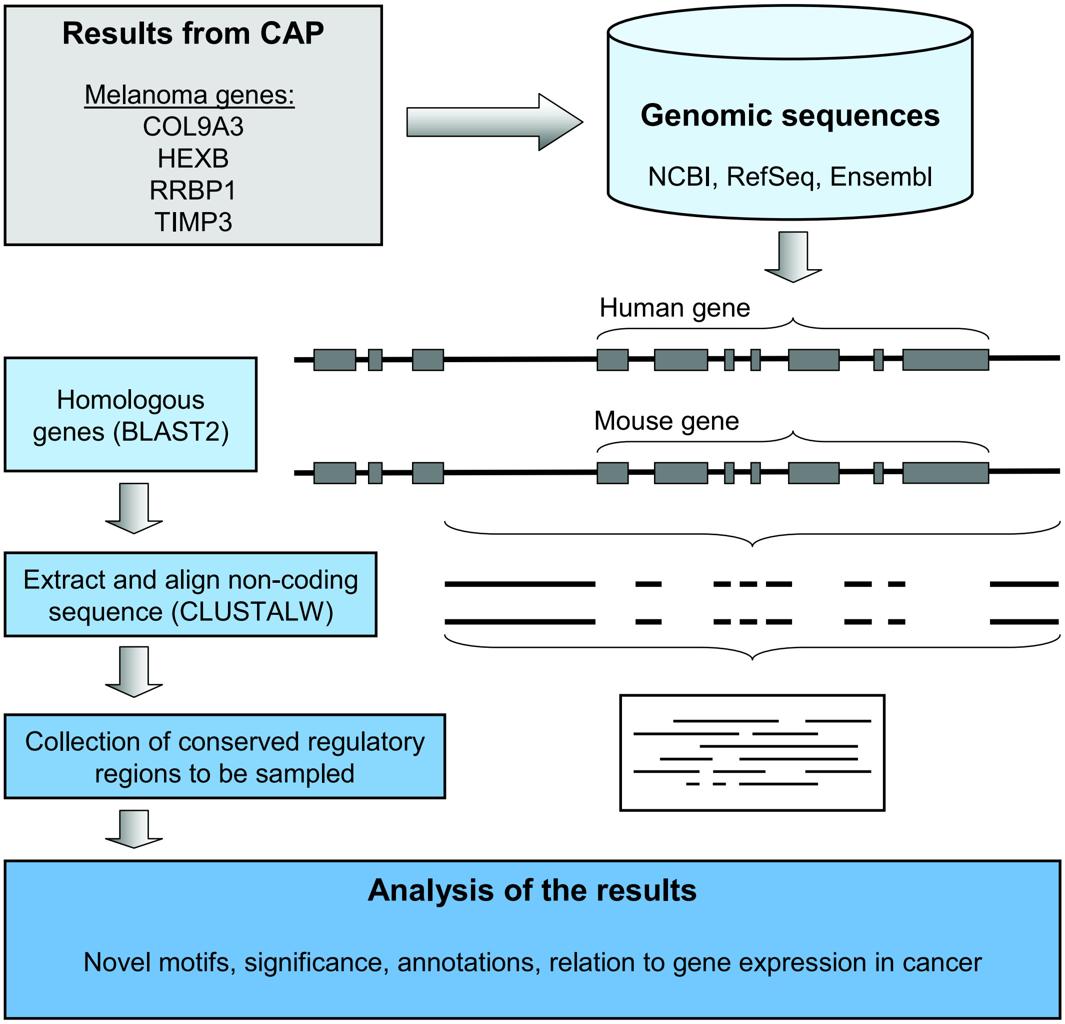

Results from CAP
Melanoma genes:
COL9A3
HEXB
RRBP1
TIMP3
Genomic sequences
NCBI, RefSeq, Ensembl
Human gene
Homologous genes (BLAST2)
Mouse gene
Extract and align non-coding sequence (CLUSTALW)
Collection of conserved regulatory regions to be sampled
Analysis of the results
Novel motifs, significance, annotations, relation to gene expression in cancer

Figure 5.1: The results from the CAP study define the set of *H. sapiens and M. musculus* genomic sequences and annotations to be extracted from sequence databases. A collection of conserved regulatory sequences is obtained using phylogenetic footprinting i.e. aligning sequences and extracting conserved fragments. The Gibbs sampling algorithm and the MEME program are used for discovering over-represented and statistically significant motifs. The significance of the novel motifs is further evaluated, each motif is compared to known reference motifs in the JASPAR database [217], and analyzed with respect to the relation to gene expression in cancer and putative underlying transcription logic.

Table 5.1: Data sources and tools with their corresponding URLs that were used in the analysis.

Data source or tool	**URL**
BLAST2	http://genopole.toulouse.inra.fr/blast/wblast2.html
Consite	http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite
Ensemble Genome Browser	http://www.ensembl.org
Genecards	http://bioinfo.weizmann.ac.il/cards/index.shtml
JASPAR	http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl
MEME	http://meme.sdsc.edu/meme/website/meme.html
NCBI	http://www.ncbi.nlm.nih.gov
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq

Table 5.2: The top scoring human reference motif in JASPAR for each predicted motif.

Motif name	**TF name**	**TF class**	**Score**	**Similarity (%)**
col9a3_motif	SOX-9	HMG	7.25	45.3
hexb_motif	AP2α	AP2	8.54	53.4
rrbp1_motif	PPARγ	nuclear receptor	9.74	48.7
TIMP3_motif	USF	bHLH-Zip	5.18	37.1
all_motif	IRF-2	TRP-cluster	11.81	65.6

over-represented motifs, and verification against a reference database. Additionally, literature investigation of the regulation of the genes and functional annotation of the gene products was conducted in order to explore their relatedness to cancer.

Two types of motifs could be identified: the motifs that are found in individual upstream sequences only, and the shared motifs that are found in several of the upstream sequences. The best individual motif for each gene (nomenclature: lower_case[gene name]_motif) and the best shared motif (all_motif) is presented in the following, and shown in Fig. 5.2. The best scoring motifs from the JASPAR reference collection are listed for each gene in Table 5.2. An example is the col9a3_motif, which is similar (matches to human TFs are given priority) to the reference motif SOX-9. Database hits for the TFs were identified for each individual gene. The list of obtained reference motifs was used for further investigation of previously reported involvement in or regulation of cancer-related genes by conducting manual literature search. The reference motifs in JASPAR and the motifs obtained through sampling are not likely to share a 100% similarity, simply due to the scarce amount of sequences used in this study. However, general sequence recognition patterns should match and can give clues about the type or family of TF that could be responsible for the regulation.

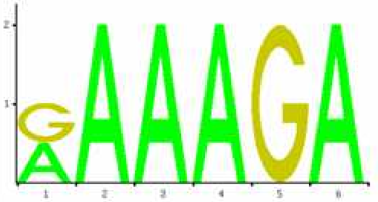
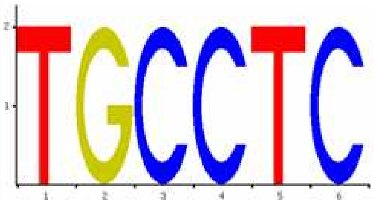
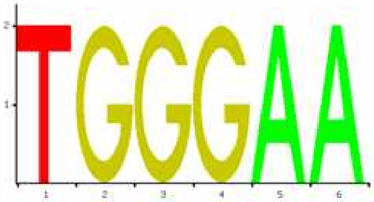
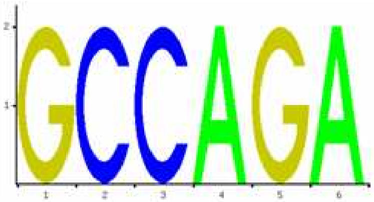
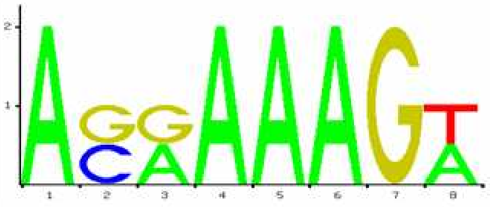
Motif ID	Motif logo	Length	# Occ.
col9a3_motif		6	8
hexb_motif		6	5
rrbp1_motif		6	8
timp3_motif		6	5
all_motif		8	15

Figure 5.2: The detected motifs in the conserved blocks of the regulatory regions upstream of the genes (indicated by the motif names), sequence logo, motif lengths, and number of occurrences in the respective upstream sequences.

Here follows a detailed discussion of some of the matching TF reference motifs and their putative role in the regulatory control of these melanoma-related genes.

SOX The SOX-gene family [64] represents an ancient group of TFs involved in numerous developmental processes and sex determination in vertebrates. SOX proteins are characterized by a conserved high mobility group (HMG)-box domain, which is responsible for DNA binding and bending.

AP2 The AP2 [13, 250] TF family is a set of developmentally regulated, retinoic acid inducible genes composed of four related factors, AP2α, β, γ, and δ. The AP2 factors orchestrate a variety of cellular processes including apoptosis, cell growth, and tissue differentiation during embryogenesis [155], which all are important aspects of melanoma cancer development [250]. Loss of AP2 function has been experimentally verified in metastatic melanoma cells [12].

PPAR Peroxisome proliferator-activated receptors (PPARs) [181] belong to the nuclear receptor superfamily that includes receptors for steroids, thyroid hormone, vitamin D, and retinoic acid. Three receptor subtypes of PPARs, termed α, β, and γ have been identified. These function as ligand-activated TFs and control the expression of genes implicated in extra- and intracellular lipid metabolism.

USF The USF (upstream TF) [85] encodes a member of the basic helix-loop-helix leucine zipper (bHLH-Zip) family, and is known for its involvement in development and differentiation [107].

IRF Shared motif: Interferon regulatory factor-1 (IRF-1) and IRF-2 are nuclear TFs that respond to interferon gamma (INFγ) [164]. INFγ is a cytokine that is produced by T-cells and natural-killer cells and has a variety of immunological effects.

Transcriptional control clearly plays an important role in the complex events underlying the initialization of tumor growth and metastasis of human cancer. Over-expression is a sign of a disturbance or dysfunction of the underlying regulatory network. The regulatory mechanisms controlling the genes, which were shown to be over-expressed and responsible for triggering an immune response in melanoma are addressed in this study. On the sequence level it is possible to study the exact mechanisms, by identifying which TFs are involved that could cause over-expression of the individual genes. The resulting motifs in combination with a manual literature search show that there are putative functional TFBSs, in each of the upstream regions of the melanoma-related genes. Especially, the AP2 and the INF TFs have previously been shown to regulate gene expression in melanoma tumors and to play pivotal roles in growth of melanoma tumors [12, 13]. Furthermore, the identification of a motif, similar to that of the IRF TFBS, supports the evidence that the genes induce an immune response in melanoma [164] (which was found using CAP). This study could be extended and was primarily restricted by the small number of melanoma-related genes and the high complexity of human genomic DNA sequences. The results supports previously reported experimental evidence and illustrates how sequence-based analysis can be conducted on a small scale on interesting experimental data.

5.2 Integrated analysis of biologically linked clusters of genes in yeast

5.2.1 Introduction

There is a clear need for flexible tools allowing for fast analysis of the regulatory regions of many different sets of genes. Such systems can help in finding answers to immediate questions posed by the experiments. In order to address this need, the yeast regulatory sequence analysis system (YRSA) [218] was developed.

Baker's yeast (*S. cerevisiae*) is a popular model organism used for studying complex regulatory networks. Compared to the human and mouse genome, the yeast genome contains significantly less genes and the upstream sequences are typically shorter than 1000 bp. Nevertheless, gene regulation in yeast is far more complex than that in bacteria and several networks and cascades in yeast are as complex as in human. Cooperatively functioning genes are commonly regulated in a similar manner ensuring complete sets of interacting proteins are present at the same time. As yeast is one of the commonly used model organism, immense amounts of experimental data has been produced using microarray expression profiling [212, 235], ChIP assays [117, 209, 229], phenotypic screens [89], text analysis [238], and protein-protein interactions [116, 258]. Key advances have been made in defining gene networks and their common transcriptional control mechanism [110, 111, 193].

The vast amount of experimental data holds an enormous potential for governing insights into the coordinated transcription of biologically linked genes in transcriptional networks. Integration of the required steps in the analysis process is a major challenge. The aim of an exhaustive promoter analysis is to decipher common characteristics and link these to the coordinated expression of gene networks. Effective means of comparing and assessing motifs combined with current methodologies is required for expert analysis. The developed system YRSA is described in detail and its usefulness is illustrated through a set of case studies. The results both re-affirm previously published work and reveal interesting findings as applied to novel data. The main finding is the hypothesis that the well-known cell cycle-regulated response element MCB may regulate the induced expression of a set of genes following cellular exposure to DNA-damaging agents. YRSA is available at http://YRSA.cgb.ki.se/ and further details has been published in [218].

5.2.2 Materials and methods

Promoter sequences

Sets of promoter sequences for analysis can either be uploaded or selected from a pre-defined set of 1,000 bp upstream sequences for 6,255 *S. cerevisiae* open reading frames (ORFs). To narrow the search space, users may indicate sub-segments for analysis. The promoter sequences were retrieved from the SCPD database (http://cgsigma.cshl.org/jian/) [290].

Motif representation

TF binding preferences were modeled using PSSMs, which describe the counts of base occurrences in every position of the binding site. For visual representation of TF binding profiles, sequence logos [222] are used.

Pattern discovery algorithm

For pattern discovery, the Gibbs motif sampler by Lawrence *et al.* [146] is utilized. This software has been continuously developed since the original publication. For the systematic study of a reference collection of gene sets, the estimated number of sites was set to one site/sequence, using 10-bp pattern width and a filter to remove low complexity regions.

Model collection

The YRSA database contains a collection of non-redundant PSSMs for 38 yeast TFs. Sequence data used for the construction of models was collected from the published literature and high-quality data repositories produced by leading research groups, including SCPD [290] and the ACE collection [110, 212]. Sequences were aligned in order to obtain PSSMs, using the Gibbs Motif Sampler. The collection of models is stored in a relational database system (MySQL) and is available online (http://yrsa.cgb.ki.se/matrixlist.html). Pattern searching and scoring PSSMs can be used to quantitatively score any sequence for its potential to serve as a TFBS for the indicated TF. By sliding PSSMs along promoter sequences, high scoring frames (hits) may be identified and classified as putative TFBS. Since the matrix models can be of varying length, their respective score ranges are unique. To be able to compare putative hits for different PSSMs, the hit scores (H) are first converted into a unit scale (normalized hit score, H_N) given by:

$$H_N = \tfrac{100(H-m)}{M-m}$$

where m is the minimum and M the maximal possible hit score for the subject matrix.

Object-oriented TFBS modules

We used the Perl TFBS [154] extension to BioPerl [237] (http://www.bioperl.org). TFBS is an object-oriented Perl module with C extensions that allows for promoter scanning, model creation, and model handling, using Gibbs sampling, relational database systems, and other tasks involved in promoter analysis. As an expansion of TFBS, we added modules for manipulating sets of sequences and visual interpretation of site distributions, as well as an interface to pattern comparison algorithms.

Comparison of binding profiles

For comparison of binding profiles, a novel quantitative comparison algorithm was applied. Matrix models in normalized formats are compared using a novel scoring function. The scoring function S evaluates the similarity between two columns x and y of the two matrices X and Y.

$$S = 2 - \sum_{b \varepsilon [A,C,G,T]} (x_b - y_b)^2$$

The scoring function gives a score between a maximum of 2 (total identity) and minimum of 0 for each position in a given alignment. The total score S_{tot} of a given alignment is the sum of call column pairs i.

$$S_{tot} = \sum_{b \varepsilon [i]} S_i$$

Finally, the normalized total score N is obtained by dividing S_{tot} with the potential maximum total score according to:

$$N = \frac{S_{tot}}{2w}$$

where w the shortest width of the two aligned motifs.

The motif alignment is obtained using a Needleman-Wunsch algorithm [184], which allows for the opening of a maximum of one gap in the alignment. This addresses situations in which TFs bind as heterodimers with variable spacing between half-sites, which for instance is the case in leucine zipper structures [131], the δEF1 ZF [208], certain homeodomain proteins [286], and nuclear receptors [211]. Based on observations in a study with metazoan binding profiles, gap opening and extension penalties were set to -3 and -0.01, respectively, for the test cases. This allows for an optimal semi-global alignment of the matrix models, given the scoring function S and gap penalties defined above.

5.2.3 Results

Performance

Over-represented sequence patterns were obtained from the regulatory sequences of associated genes using the YRSA system. A series of tests were conducted in order to assess the reliability and the performance of the predicted TFBS motifs.

Extension test for sensitivity

The limitation of the pattern finding was assessed by obtaining known binding sites for three TFs annotated in the SCPD database, listed in Table 5.3. The individual binsing sites for each TF were extended by flanking sequence from their natural promoters. The length of the flanking sequence was gradually increased from 25 to 225 bp in both directions. Each study was repeated 30 times and compared against a control set of randomly chosen promoter sequences. The observed patterns were compared to the reference models in the YRSA database to assess the average comparison score and rank obtained by the motif comparison algorithm, these results are plotted in Fig. 5.3 a and b. The sensitivity of the motif finding algorithm decreases as the flanking sequence is progressively extended. Pattern detection begins to fail when the signal-to-noise ratio becomes low. The precise signal detection limit is unique to each TF. The observed results are consistent with previously published reports, and the drop in score at the lower bound is presumed to originate from the inclusion of false sites in the profile.

Analysis of a reference set for specificity

The performance of YRSA was tested on a reference collection of gene sets to determine the specificity of the detected patterns. A data set collected by Lee *et al.* [150] from published literature contains sets of genes known to respond to specific TFs (regulons). The precise location of the motif is unknown in many cases and there is no exclusive evidence that each listed promoter contains a motif. Thus, this data set presents a real-world challenge to YRSA. Putatively co-regulated genes for which a binding profile was available in the YRSA database

Table 5.3: TFBSs and promoters used in the extension test.

Transcription factor	Gene ID	Binding site sequence	Start	Stop
REB1 (model ID: MY0021)	YDL164C	TTACCCGGAT	-351	-342
	YDR007 W	CGCTCACCCG	-322	-313
	YDR050C	AGATTACCCG	-401	-392
	YDR146C	TCCGGGCAA	-361	-353
	YER086 W	CGGGTAGCAA	-186	-177
	YFL039C	CTGTCACCCGGCC	-410	-398
	YGL026C	TTATTACCCG	-169	-160
	YNL216 W	CCGCTACCCG	-232	-223
	YOL004 W	TTACCCGTGT	-372	-363
	YOL006C	TCCGGGTAA	-287	-279
	YPL231 W	TTACCCG	-296	-290
MCB (model ID: MY0013)	YDL102 W	ACGCGT	-166	-161
	YDL164C	ACGCGA	-133	-128
	YJL194 W	ACGCGA	-216	-211
	YNL102 W	ACGCGT	-208	-203
	YOR074C	ACGCGT	-158	-153
CAR_REP (model ID: MY0039)	YAL005C	TCGGCGGCA	-246	-238
	YDR256C	AGCCGCGCA	-249	-241
	YGR088 W	CTGCAGGCT	-197	-189
	YGR254 W	AGCCACCTC	-219	-211
	YGR264C	TGTTAGCCGCCGA	-248	-232
	YHR018C	GTGGTGGTT	-229	-221
	YML054C	AACCGCCAA	-167	-159
	YMR108 W	AGCCGCCGG	-483	-475
	YOL006C	AGCCGCCGA	-447	-439
	YPL111 W	AGCCGCCGA	-156	-148

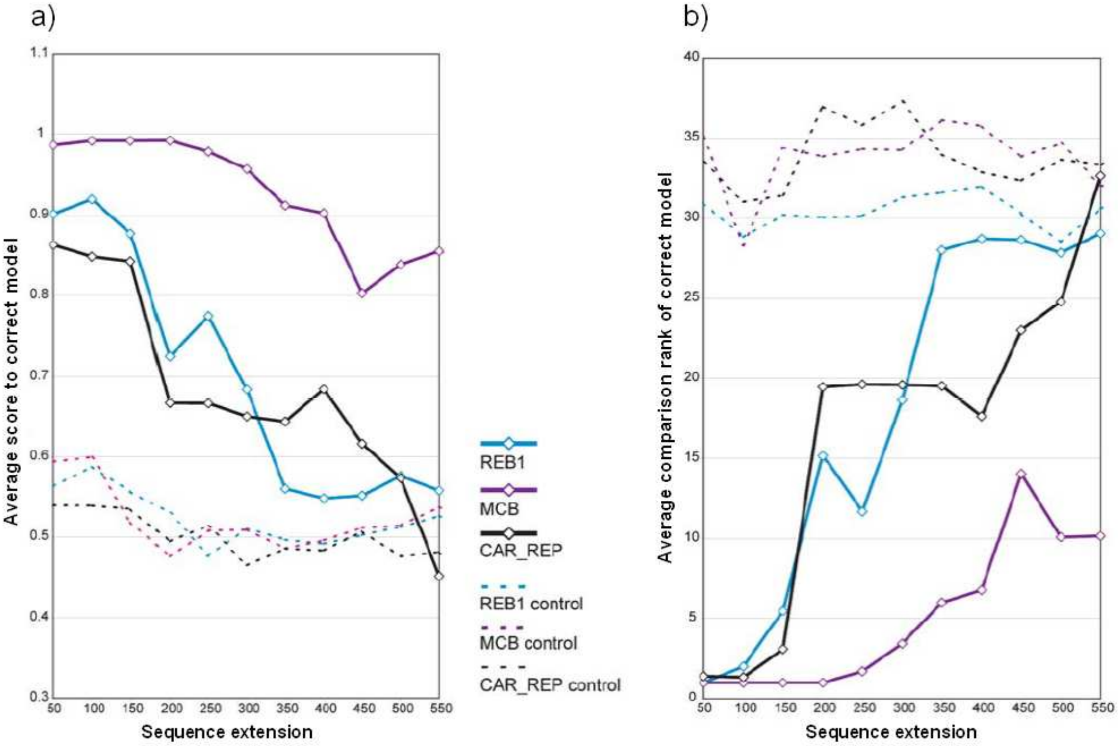


Figure 5.3: In the extension test for sensitivity, the motifs for three representative TFs (REB1, MCB, and CAR_REP) were obtained using the YRSA system and an increasing flanking sequence around the annotated motif. These motifs were compared to the known motifs in the YRSA database and the average score (a) and average rank (b) are plotted against the extended flanking sequence.

were grouped into *regulons*. The 16 selected regulons contain a varying number of sequences (from 5 to 28 genes), for which the promoter segments -500 to -1 were obtained. The test is not comprehensive (detailed results can be found in [218]), but they indicate that promising motifs can be obtained even for larger and less tightly defined gene sets.

Analysis of experimental data

The extension test and the analysis of the reference collection showed that the YRSA system is applicable in the motif discovery process of sequences mediating transcription of yeast genes. YRSA has been applied to a series of studies presenting an increasing level of difficulty of motif detection.

Case study 1: Analysis of PDR1/3-regulated genes

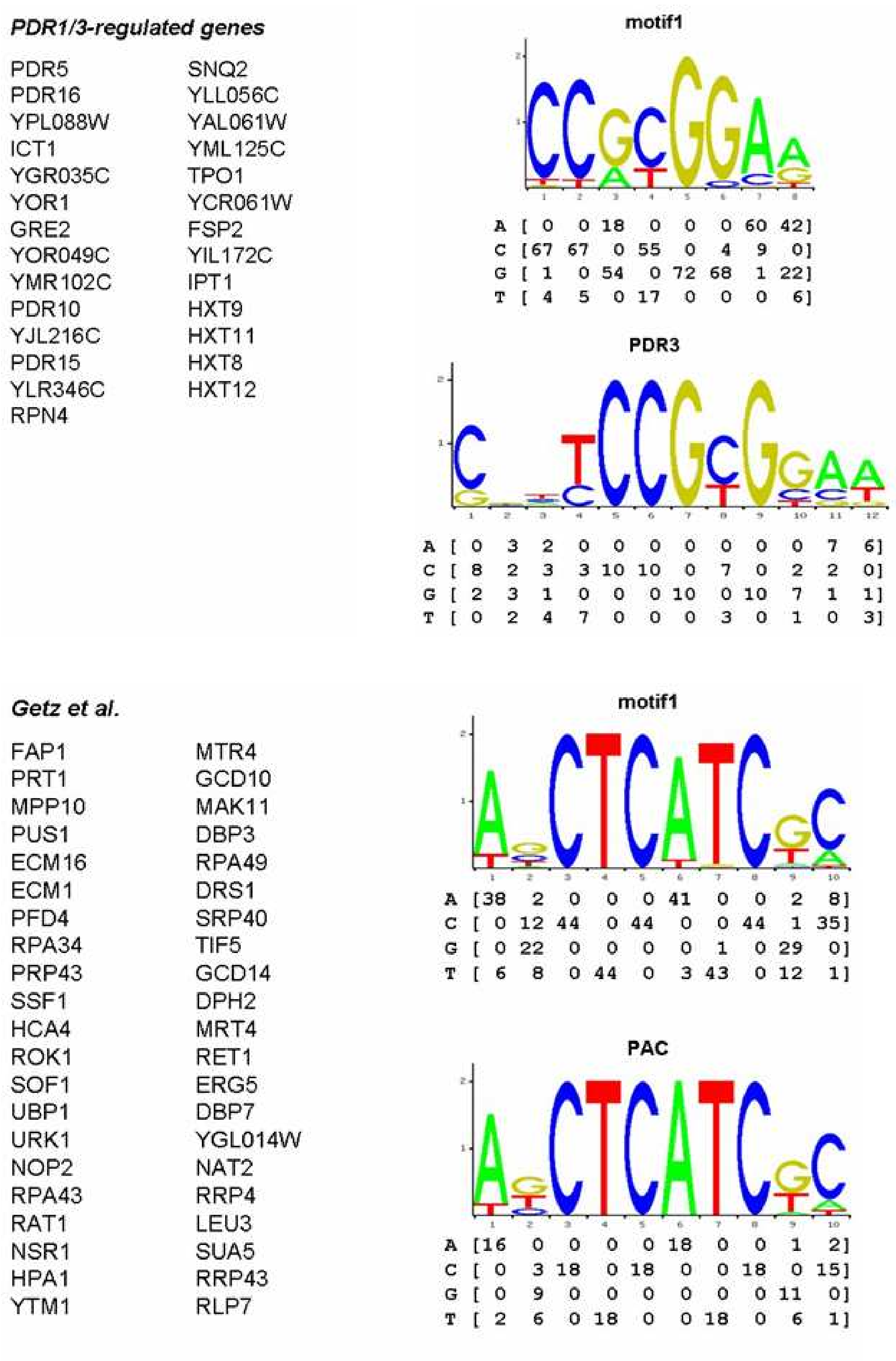
The first application of YRSA is on a regulon defined by multiple microarray studies of genes responsive to the TF PDR1 [61]. PDR1 is closely related to PDR3, published experimental data lead to the definition of a common set of target genes. The PDR1/3 regulon contains 27 genes listed in Fig. 5.4. YRSA can detect a motif (motif1 in Fig. 5.4 a), which is highly similar to the annotated motif for PDR3 in the YRSA database. Potential circularity problems are avoided since only three of the 27 input sequences were used for constructing this motif [110, 212]. The specificity of the PDR1/3 motif appears to be high since it could be detected in long promoter sequences (700 bp, data not shown).

Case study 2: Genes activated by an unknown TF following release from cell cycle arrest

Getz et. al. [88] applied a novel super-paramagnetic clustering algorithm on the well-annotated yeast cell microarray data set by Spellman *et al.* [235]. Focusing on finding sets of putatively co-regulated genes with ascribed function, an interesting regulon (consisting of 42 genes listed in Fig. 5.4) related to metabolism was identified. Promoter analysis revealed a highly conserved motif (consensus sequence ANCTCATCGC), and the authors concluded that it most likely is the binding site for an unknown TF. YRSA finds an over-represented motif, which is highly similar to the PAC motif in the YRSA database shown in Fig. 5.4.

Case study 3: Analysis of DNA-damage response genes

In this third study YRSA is applied to a novel data set, for which no promoter analysis has been reported. The DNA damage repair system is crucial for avoiding fatal mutations in all



PDR1/3-regulated genes

PDR5
PDR16
YPL088W
ICT1
YGR035C
YOR1
GRE2
YOR049C
YMR102C
PDR10
YJL216C
PDR15
YLR346C
RPN4
SNQ2
YLL056C
YAL061W
YML125C
TPO1
YCR061W
FSP2
YIL172C
IPT1
HXT9
HXT11
HXT8
HXT12

motif1

```
A [ 0  0 18  0  0  0 60 42]
C [67 67  0 55  0  4  9  0]
G [ 1  0 54  0 72 68  1 22]
T [ 4  5  0 17  0  0  0  6]
```

PDR3

```
A [ 0  3  2  0  0  0  0  0  0  0  7  6]
C [ 8  2  3  3 10 10  0  7  0  2  2  0]
G [ 2  3  1  0  0  0 10  0 10  7  1  1]
T [ 0  2  4  7  0  0  0  3  0  1  0  3]
```

Getz et al.

FAP1
PRT1
MPP10
PUS1
ECM16
ECM1
PFD4
RPA34
PRP43
SSF1
HCA4
ROK1
SOF1
UBP1
URK1
NOP2
RPA43
RAT1
NSR1
HPA1
YTM1
MTR4
GCD10
MAK11
DBP3
RPA49
DRS1
SRP40
TIF5
GCD14
DPH2
MRT4
RET1
ERG5
DBP7
YGL014W
NAT2
RRP4
LEU3
SUA5
RRP43
RLP7

motif1

```
A [38  2  0  0  0 41  0  0  2  8]
C [ 0 12 44  0 44  0  0 44  1 35]
G [ 0 22  0  0  0  0  1  0 29  0]
T [ 6  8  0 44  0  3 43  0 12  1]
```

PAC

```
A [16  0  0  0  0 18  0  0  1  2]
C [ 0  3 18  0 18  0  0 18  0 15]
G [ 0  9  0  0  0  0  0  0 11  0]
T [ 2  6  0 18  0  0 18  0  6  1]
```

Figure 5.4: The genes in the PDR1/3 regulon (top) are listed next to the motif (motif1) detected using YRSA. This motif is highly similar to the PDR3 reference motif in the YRSA database. Another regulon by Getz *et al.* [88] (bottom) was used in the second study. YRSA identified one motif, which is highly similar to the annotated PAC binsing site [110, 212].

organisms. Gasch *et al.* [87] identified a set of genes induced by methyl-methane sulfonate (MMS) or radiation. Additional data was used to refine the regulon and the authors reported a condensed cluster of nine putatively co-regulated genes listed in Fig. 5.5. Several genes were known from earlier studies to be connected to DNA repair mechanisms, whereas other genes in the regulon have no annotated function. Little is know about the repair mechanism, with the exception of the Mec1 response [69] and the coupled Crt1 repressor [109]. Gasch *et al.* noted that only a subset of the DNA-damage response set of genes overlaps the Mec1 response, suggesting the existence of additional control mechanisms.

Applying YRSA to these MMS-induced DNA repair genes reveals a motif similar ot the MCB (MluI cell cycle box [110, 212]), which is bound by the MBP1 protein and the cofactor SWI6, could be identified, see Fig. 5.5. Based on this observation, we suggest that the DNA damage response is partially regulated by MBP1 or a TF with highly similar binding specificity. Interestingly, it is known that the DNA-damage and replication pathways in yeast activate checkpoints at four stages of the cell cycle. The MCB element partially controls the entry from G1 phase into S1, where DNA replication takes place [276]). The role of MCB sites in blocking DNA replication and the presence of a suitable target site in the promoters of DNA-damage responsive genes (the matching TFBSs in each of the nine promoters are graphically illustrated in Fig. 5.5) is consistent with the biological processes.

5.2.4 Discussion

Pattern discovery methods can be successfully applied to a set of transcriptionally co-regulated yeast genes for the identification of the TFBSs in the regulatory regions. The presented YRSA system shows a successful integration facilitating studies of co-regulated genes. The integrated analysis tools, including a Gibbs sampling method for motif detection and a novel algorithm for motif comparison, allow for rapid assessment of predicted regulatory controls of gene clusters. The large-scale systematic assessment of YRSA presented here, shows its applicability to biological data.

Applying YRSA to experimental data resulted in a hypothesis consistent with what has been previously observed. The MCB regulatory element or an element of a similar TF plays a mediating role in the DNA damage response. Compared to previously published work conducted on individual genes of DNA damage response, this study presents an analysis of the full set of putative target genes confirming earlier work. TFs with similar binding specificity and DNA-binding domain as the MCB-associated protein MBP1, are SWI4, SOK2,

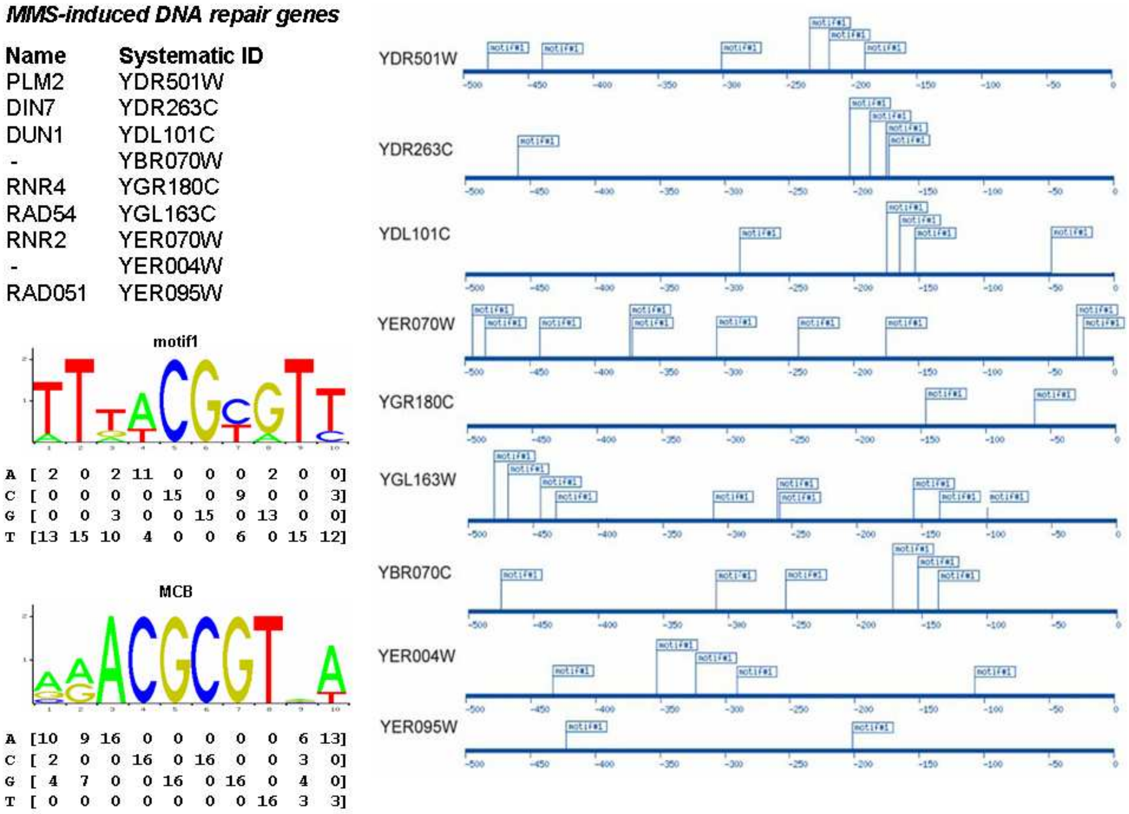


Figure 5.5: The nine yeast MMS-induced DNA repair genes are listed to the left, below the detected motif and its best match in the YRSA database (MCB). The distribution of the potential matches in the promoter sequences (-500 to -1) are illustrated to the right.

Table 5.4: Integrated web-based systems for yeast promoter analysis.

System	Description	URL
ACE	AlignAce (Pattern recognition) CompareAce (Pattern comparison)	http://atlas.med.harvard.edu/ http://atlas.med.harvard.edu/cgi-bin/compareace.pl
RSAT	Pattern recognition and scans	http://embnet.cifn.unam.mx/rsa-tools/
INCLUSive	Pattern recognition and scans	http://www.esat.kuleuven.ac.be/∼dna/BioI/Software.html

and PHD1.

Detecting TFBSs is a true challenge in bioinformatics. The short and often variable functional binding sites must be distinguished against an often noisy promoter background. The performance assessment shows that the length of yeast promoters is at the upper limit for the motif detection method. Further tests show the potential of the YRSA system on experimental data, likely to contain noise, for analyzing the TFs mediating transcription of a set of biologically linked gene clusters.

YRSA integrates pattern recognition, detection of sites in promoter sequences, and the classification of TFs likely to act through a newly detected pattern. Each of these components has been the focus of individual projects. Three previous reports have defined efforts to couple at least two of the steps: ACE [110], RSAT [261], and INCLUSive [252], which are listed in Table 5.4. The RSAT and INCLUSive services partially overlap the functionality of YRSA in pattern recognition, but should be viewed as complementary methods. Both allow users to scan multiple genomes (including yeast) for sites matching a binding profile and to retrieve sequences flanking predicted sites. The ACE suite of tools is the only set to include a complete set of pattern recognition and classification components comparable to those within YRSA. However, the tools generated for specific applications are not integrated and lack many features for user control. The comparison algorithm is substantially different and highlights the novelty of YRSA. ACE uses a comparison algorithm based in large part on a method for comparing ungapped protein alignments [110, 196]. YRSA introduces a novel scoring function that allows for gaps. Since there is no limit on the motif widths, an unwieldy constraint of a specific number of significant positions within binding sites is avoided.

The distinguishing component of YRSA is the underlying database of reference motifs. These were obtained from literature and as new technologies emerge and research projects progress, the body of experimental evidence for both known and unknown TFBSs grows. It is crucial to keep the reference collection up-to-date by editing and adding new data. Enabling user-specified choice of further pattern detection methods is a clear advantage to YRSA. Not included in this study, but an obvious future direction is to incorporate phylogenetic footprinting. The promising advantage of utilizing information about conservation across species is of definite advantage for identifying functional genomic sequences.

5.3 Regulatory networks determining stem cell fate in plant

5.3.1 Introduction

The beautiful shape and color patterns of flowers attract almost everybody, including developmental biologists. The genome of *A. thaliana* (also known as mouse-ear cress) was the first plant genome to be fully sequenced and is commonly used as a model organism for studying plant development. The regulatory regions of plant genes tend to be more compact than those of animal genes. However, plants have as many TFs and the complexity of the control mechanisms is as high, which provides an excellent model system to study transcriptional programs and the developmental stages of multicellular organisms. A further difference between plants and animals, is that plants develop continuously (i.e. form new organs postembryonically) presenting all stages of the life cycle within the repeating units of an individual (e.g. leaves). The cellular basis for this mode of development is the stem cells that are located in the apical meristem of the shoot and root, which are the growing parts of the plant. The number of stem cells in the pool is small, typically six to nine, and under tight transcriptional regulation. Some key TFs have been identified, however, exactly how these mediate and control their target genes still remains to be elucidated.

The aerial tissues (leaf and flower primordia) of plants are generated from the plant apex - a small cone-shaped region called the shoot apical meristem (SAM) [239] that plays an important role in plant development [271]. An important control circuit, modulating the size of the SAM, is a feed-back loop between the WUSCHEL (WUS[1]) and the CLAVATA3 (CLV3) gene expression [230]. The expression of WUS and CLV3 has been shown to be spatially separated (confined to different expression domains), which indicates that signal transduction is necessary for communication. WUS expression is activated in and restricted to the cells in the organizing center (OC) [174], which is directly situated below the SAM, illustrated in Fig. 5.7. CLV3, on the other hand, is expressed by the cells in the SAM. The WUS gene encodes for a nuclear-localized homeodomain (HD) TF that promotes initiation and maintenance of the stem cells in the apex [145, 174], by inducing CLV3 expression in the SAM. Transmission of the CLV3 (short peptide) signal is carried out through lateral movement [153] and binding to the CLV1/CLV2 kinase receptor complex, which causes restriction of WUS activity to the OC [32]. A 57-bp region, in the WUS upstream sequence (approximately at

[1] The gene name is abbreviated using capitals (WUS), whereas a specimen with a knock-out mutation is abbreviated in italics and lowercase (*wus*).

Figure 5.6: The genome of *A. thaliana* (also known as the mouse-ear cress) was the first plant genome to be fully sequenced. This flower has become a model organism and is commonly used for studying developmental biology in plant. The picture was obtained form http://www.weigelworld.org/research/gallery/.

-550) appears to be necessary and sufficient for WUS expression in the SAM, was shown to contain two regulatory elements (RE1 and RE2). RE1 (-566 to -557) has the sequence *TTGAGAAGAG* and overlaps with three bp of the HD-Zip consensus. RE2 (-546 to -541) has the sequence *TGAAAA*. The position in the upstream regions seems to be crucial to the function for the TFBSs, perhaps due to a favorable chromatin structure [35]. In *wus* mutants, stem cells differentiate prematurely and, therefore, the SAM is greatly reduced, whereas in *clv3* mutants, stem cells proliferate abnormally leading to a enlarged SAM. Inducible overexpression of WUS or CLV3 can cause the reverse effects.

There is a second pathway that has been shown to control the stem cell population. The SHOOT-MERISTEMLESS (STM) gene encodes a HD TF, which is a second mediator amplifying the stem cell daughter cells in the SAM [162]. The SAM is absent in *stm* mutants. STM (in a similar way as WUS) acts through a feed-back loop [265], but has a different set of target genes. STM and WUS seem to require each other for the correct regulation of the stem cell population [15, 174].

Meristem plasticity requires the ZWILLE (ZLL) gene. In *zll* mutant embryos, the apical cells are defect causing a variability of the meristem's size and function. Therefore, two phenotypes can be observed in *zll* mutant seedlings: one lacks the SAM and in the other one the SAM develops into a pin-like structure.

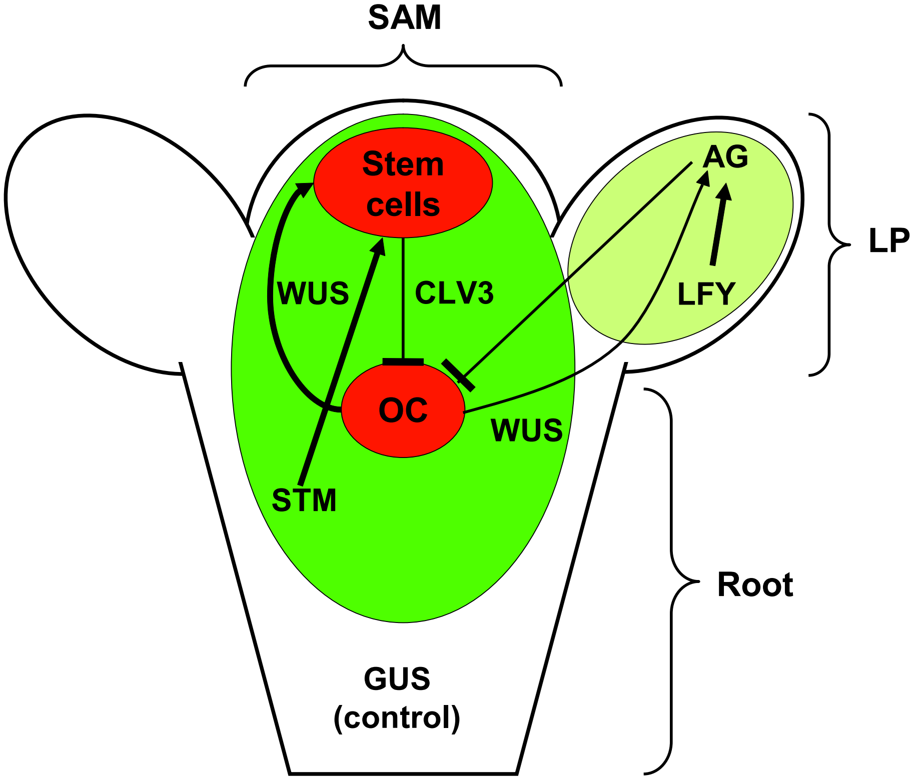


Figure 5.7: An illustration of the current theoretical model of gene regulation in the aerial organs, the shoot apical meristem (SAM) and the leaf primordia (LP), of *A. thaliana*. First, the negative feed-back loop in the SAM (WUS/CLV3). The cells in the organizing center (OC) express WUS, which induces stem cell identity in a few cells above the OC. The stem cells express CLV3, which in turn represses WUS activity in the OC. The stem cell population is further and independently regulated by STM. Second, the negative feed-back loop (WUS/AG) connecting the gene expression in the SAM and the LP is also illustrated. The expression of WUS in the SAM induces AG expression in LP, which in turn represses the activity of WUS. The expression of AG is further regulated by LFY.

In contrast, the LEAFY (LFY) TF is responsible for the formation of leaf primordia (LP) and a direct regulator of the AGAMOUS (AG) and APETALA1 (AP1) genes. AG and AP1 are both floral homeotic[2] genes [29]. Hence, LFY brings about terminal differentiation of young cells in the LP. WUS and LFY cooperate to activate AG in the center of the LP, which induces the identity of stamens and carpels and limits cellular proliferation. A direct link between the control of the stem cell pool (WUS/CLV circuit) and the activation of floral homeotic gene expression in the LP, was established as it could be shown that WUS induces AG in the LP contributing to meristem growth and that AG in turn represses WUS through a negative feed-back loop [161]. However, no AG consensus [232] has been identified in the WUS promoter. WUS and LFY bind to regulatory sequences in the second intron of the AG gene, where a few additional highly conserved response elements (RE) have been identified [106]. The first RE1 has the sequence $CCAATCA$, which is similar to the binding site of a TF called Nuclear Factor-Y (NF-Y, which has three homologs A, B, and C). The sequence of RE2 is $aAGAAT$, but does not match any known consensus sequence. The $CArG$ (r being 6-7 nucleotides) motif is bound by MADS domain TFs [186] and AGL6 has been identified as a putative candidate [106].

WUS appears to play a dual role: it controls the stem cell population and activates floral homeotic genes. AG also shows this duality as it induces floral organ identity and determines the growth of the floral meristem. However, there is a fundamental difference between the two negative feed-back loops. The WUS/CLV signaling pathway occurs in the same cells, but is temporally separated, whereas the WUS/AG circuit takes place in adjacent cells simultaneously. The current regulatory model is based on interaction studies mainly through loss-of-function (LOF) and gain-of-function (GOF) experiments [77, 31, 223]. The robustness of the network is impressive, since it varies throughout the different developmental phases and has been shown to self-organize back to normal patterns after induced disturbance [207]. Diverse regulatory pathways seem to converge at a central transactivating mechanism [17]. Much data supporting the current theories exist but there are still many open questions regarding the transcriptional network and an attempt to model SAM development computationally was recently reported [124].

Since most of the known stem cell regulators in *A. thaliana* are TFs, we have set out to elucidate the regulatory network of stem cell control by computational analysis of experimental data. In order to explore the regulatory networks in the SAM and the LP, transcriptional

[2] A gene that affects embryo development by specifying the character of a body segment.

profiling experiments are employed for investigating the key TFs and their target genes. Currently we verify the microarray data by quantitative rtPCR and study the spatial expression domains and dynamics by in situ hybridization. Furthermore, we use ChIP experiments to study the interaction of the transcription factor WUS with its target genes in vivo. Common and individual target genes for WUS, CLV3, ZLL, and LFY, are identified using LOF and GOF experiments. Computational analysis of the promoter sequences is undertaken in order to gain information about gene function and the regulatory logic of stem cell control. Ultimately, the goal is to establish a comprehensive model of stem cell homeostasis with predictive power.

5.3.2 Materials and methods

The analysis process contains both experimental and computational steps and is graphically illustrated in Fig. 5.8. Putative target genes, of the key players in the regulation of stem cells and floral patterning, are identified using several transcriptional profiling experiments. Questioning the transcriptional logic with different experiments and extracting only genes that show an induced change generates a set of gene lists containing putative target genes. Analysis of these gene lists and combinations of such gene lists is conducted computationally. The promoter sequences of the putative target genes are extracted and searched for over-represented motifs that could correspond to functional TFBSs. The identified motifs are further investigated and compared to a reference collection of know motifs.

Gene expression experiments

This work is a collaboration with Wolfgang Busch and Jan Lohmann, who performed the experimental work and extraction of differentially expressed genes (briefly outlined below) at the Max Planck Institute (MPI) for Developmental biology in Tuebingen, Germany.

Pools of microscopically dissected apices were sampled at very early developmental stages, when the wild type (WT) and the mutant phenotypes just begin to deviate. The Affymetrix ATH1 array platform, which represents over 80% of the annotated genes in *A. thaliana*, was used in the microarray experiments. Affymetrics GeneChip arrays use short oligonucleotides to probe for genes in the sample. Following the image processing step further pre-processing steps need to be conducted in order to extract differentially expressed genes. First, background adjustment was performed in order to estimate the expression levels. This was done using the commonly employed gcRMA algorithm [283], which accounts for specific and non-

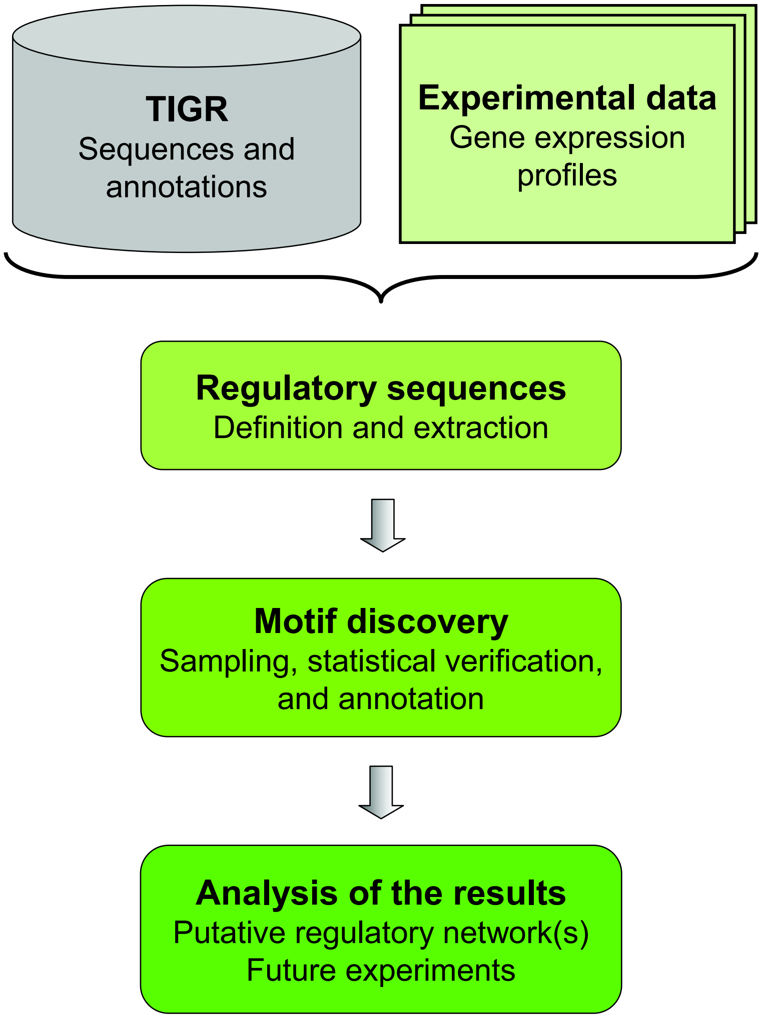

TIGR
Sequences and annotations
Experimental data
Gene expression profiles
Regulatory sequences
Definition and extraction
Motif discovery
Sampling, statistical verification, and annotation
Analysis of the results
Putative regulatory network(s)
Future experiments

Figure 5.8: The workflow of the analysis process. Experimental gene expression data define the sets of *A. thaliana* genomic sequences and annotations to be extracted from the TIGR database. The Gibbs sampling algorithm and the INCLUSive package are used for discovering over-represented and statistically significant motifs within the promoter regions of the identified target genes. The significance of the novel motifs is evaluated, compared to known reference motifs, and analyzed in respect to the experimental design and putative underlying transcription logic.

specific hybridization by regarding the difference between the pair of the perfect match probe and the mismatch probe. Secondly, differential expression was determined at the probe level by applying Logit-t [152] for normalization of the expression data. Genes showing a two-fold change in expression levels were extracted.

In order to profile the mutant plants (*wus*, *stm*, *clv3*, and *zll*), two independent experiments consisting of two replicates for each genotype were performed. Each sample consisted of 45 pooled apices dissected from plants grown in parallel. For the GOF lines, one experiment in biological triplicates was conducted. Reference samples carried inducible β-glucoronidase (GUS) as control.

Putative target genes are listed for each experimental condition. As several sets of experiments were conducted, these were used in a combinatorial fashion to extract the most likely target genes.

Sequence data

Genomic sequences of the potential target genes were obtained from the online available *A. thaliana* Database (http://www.tigr.org/tdb/e2k1/ath1/) designed and cared for by The Institute for Genomic Research (TIGR). The sequence data is stored locally in a MySQL relational database in order to enable fast and flexible extraction of non-coding regions. These sequences are defined according to variable parameters and combinations thereof, such as: upstream (or promoter) sequence length (-500, -1000, and -1500), introns, and downstream sequence. The sequence collections, which are extracted for each gene list, are used as input for the sequence sampling algorithms. The Gibbs algorithm and the INCLUSive motif finding package were used for discovering short over-represented sequence motifs, which are likely to correspond to putative functional TFBSs. The motifs are ranked according to their corresponding p-value. The p-value was assigned by generating 999 random gene lists of the same size (number of genes) as the gene list under analysis. A putative motif is used for scanning all 1000 gene lists and detecting how often the query motif occurs more often in other gene lists when compared to the original gene list. The best p-value of 0.001 is obtained if no other gene list contains the same motif more often than the analysed gene list.

5.3.3 Results

Our approach integrates expression profiling of LOF mutants and inducible GOF lines with genome-wide TF binding studies by means of computational analysis. Studying the global

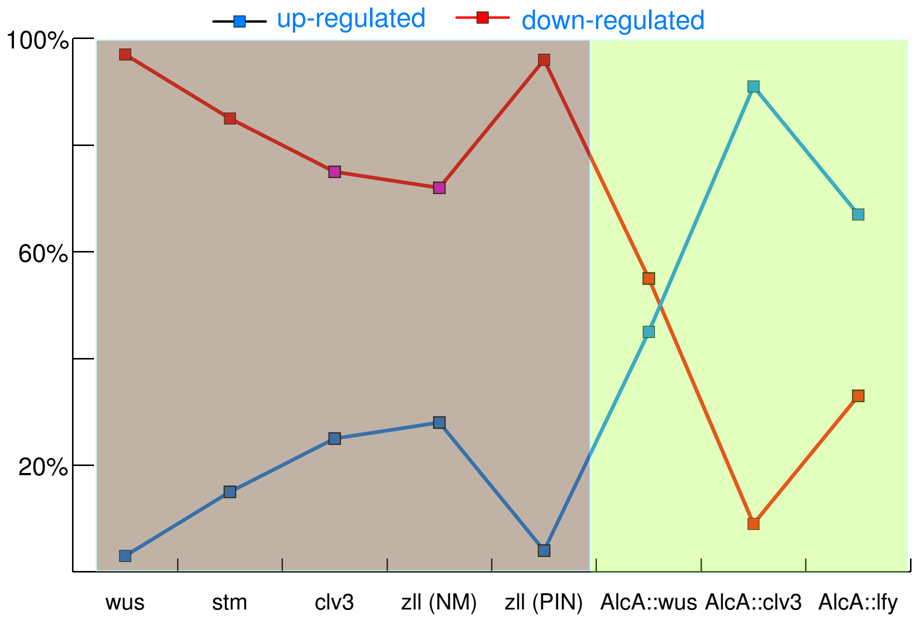


Figure 5.9: Ratio of up- and down-regulated genes in mutant and transgenic lines compared to the WT. Left (grey): mutants; right (light green): inducible over-expression lines.

effects of mutants and inducible alleles, it becomes clear that the analyzed stem cell regulators mostly cause activation of genes, see Fig. 5.9. To assess what biological processes are targeted by plant stem cell regulators, we then assigned the differentially expressed genes to functional categories (see Fig. 5.10). These annotations are defined by TAIR (The Arabidopsis Information Resource, http://www.arabidopsis.org/) and GO (Gene Ontology project, http://www.geneontology.org/). We found that many targets code for metabolic and housekeeping functions. In addition, we find a strong representation of genes involved in transcriptional control and stress response.

Using the known regulatory interactions between SAM regulators, such as the WUS/CLV3 feed-back loop, we searched for genes that follow the predicted genetic logic. For example, the expression of a true WUS target gene should not only change in the *wus* mutant, but also behave the opposite way in the *clv3* mutant. Furthermore, if it was to be a specific WUS target, it should not change in any of the other mutants, even though some display a very similar phenotype. Applying these biological criteria in addition to the bioinformatics filters described above, we were able to identify five high confidence target genes for WUS.

In order to elucidate the transcriptional logic of stem cell control, we aim to identify functionally relevant regulatory elements. To this end, we have developed a database, which allows for the automated retrieval of promoter sequences from genes identified in our profiling

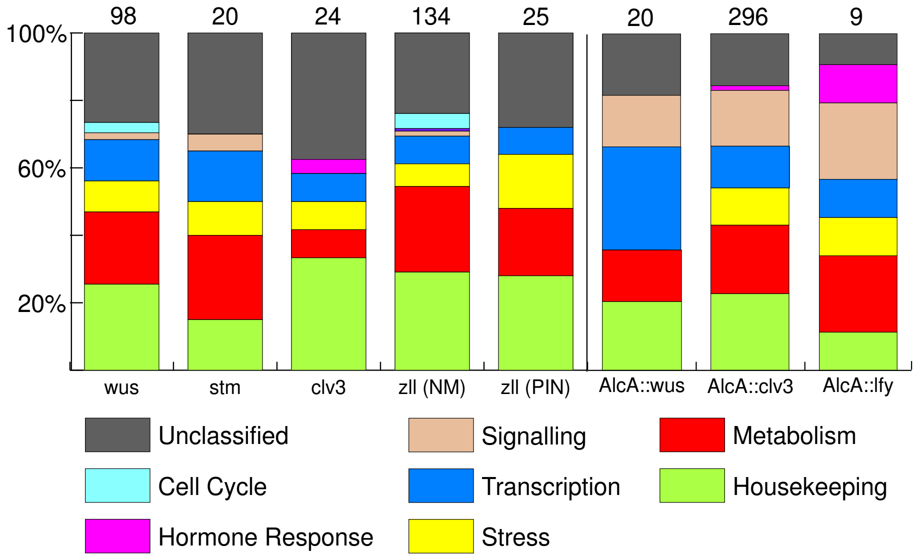


Figure 5.10: Differentially expressed genes in the different conditions assigned to global categories according to their functional annotations by TAIR and GO. The numbers above the columns indicate the number of differentially expressed genes under a certain condition compared to the WT.

experiments. Using Gibbs motif sampler and statistical tests we are currently identifying high confidence motifs from our target gene promoters. The results indicate that there are statistically over-represented motifs, both common and unique for the gene lists.

Three significant motifs are shown in Fig. 5.11. The two top ones have very good p-values, whereas the bottom one has a worse p-value but is interesting for its novelty. The distribution of each motif in the promoter sequences is graphically illustrated on the right hand side of each motif. One black pixel represents a hit and the white area along the x-axis defines the promoter regions that were extracted for the genes in the gene list. The vertical bar represents the start of a gene. This information can be helpful in determining if the motif clusters around a particular site in the upstream region. The $ACATGT$ motif (left) was found to be over-represented in several gene lists (notably in several lists that aim to identify WUS target genes) and covers a large percentage of the genes within these lists. The motif is interesting since it consists of two short palindromic sequences ACA and TGT, which also are palindromic when combined and read on the opposite strand. This binding site is characteristic of TFs in the bHLH-Zip class and it aligns with the consensus sequence $CANNTG$ (where N is any nucleotide). This structural class of TFs contains a tripartite DNA binding domain containing a basic region, a helix-loop-helix, and a leucine zipper. The TFBSs reflect the heterodimer (two short subsites) and are typically highly conserved. Proteins of this class usually play important regulatory roles in cell growth and

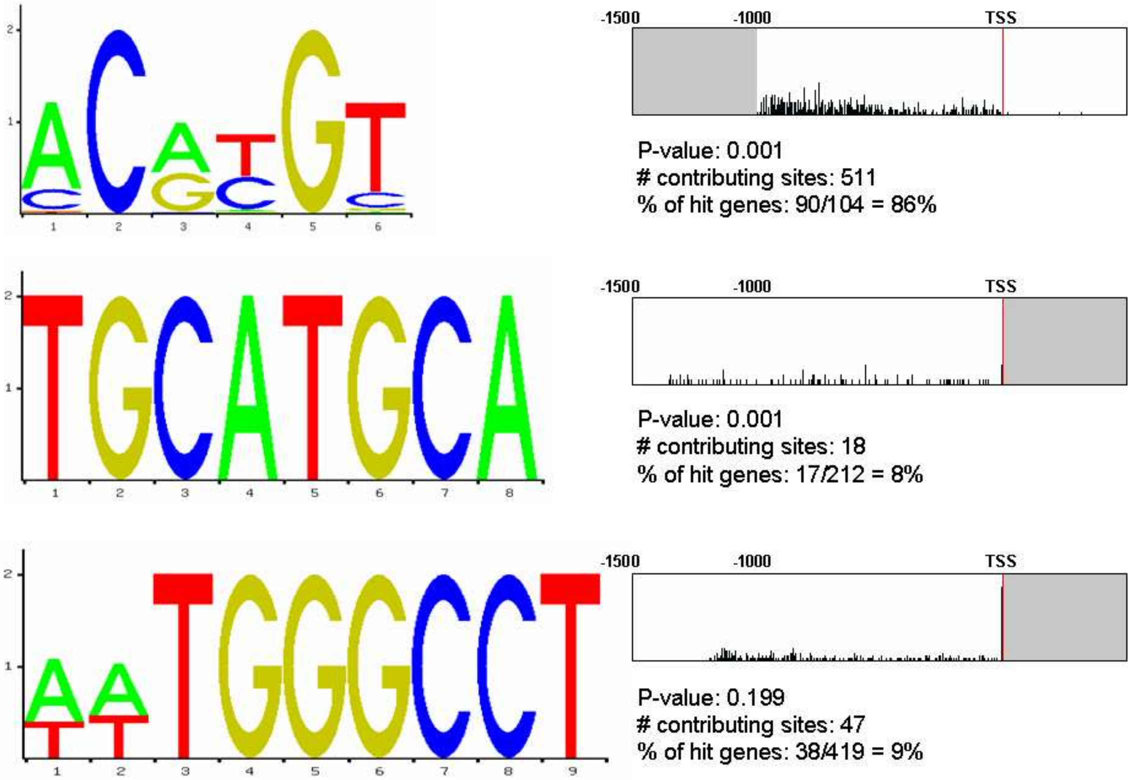


Figure 5.11: These three motifs are statistically over-represented in the promoter sequences in some of the generated gene lists. The motif at the top is highly similar to a binding site of TFs in the bHLH-Zip class. The motif in the middle is identical to a TFBS of the cAMP response element (CRE). The two motifs at the top have very low p-values and the top motif is found in 90 of the 104 genes in a particular gene list. The bottom motif is novel, i.e. does not share a significant similarity to any known motif. It is detected in a gene list where the missing link between LFY and WUS is targeted, however, the p-value is higher than for the top two motifs. The motif distribution in the promoter regions are illustrated on the right hand side. Grey areas were not used for extracting the motifs.

differentiation [107]. A potentially important function is that they regulate the rate of growth (defined as an increase in cell mass and size) that is thought to be required for cell cycle progression and cell division. The second motif $TGCATGCA$ has a low p-value and is highly conserved. It is identical to the response element of the c-AMP binding protein (CREB), which is known to be involved in regulation of cell growth [33]. The third motif $TGGGCCT$ occurs in several gene lists containing target genes that are common to LFY and WUS. This novel motif is unlike any motif in the TFBS database JASPAR [217] and could potentially be a link between the key regulators (LFY and WUS) of plant development. The motif has AT-rich flanking sequences and about 20% of the genes in a gene list has it in their upstream sequences, however, the motif is not significant.

5.3.4 Discussion

The development of multicellular organisms requires a network of highly controlled TFs. Cellular signals and transport mechanisms care for the signal transmission within one cell and between adjacent cells. The continuously developing plant *A. thaliana* is a suitable model organism for studying eukaryotic transcriptional networks, since it contains a high number of TFs and the genes have relatively short promoters compared to animals.

Experimental results show that the transcriptional control of genes with regulatory power play a pivotal role for many developmental processes. Analysis of expression profiles provides insight into the mechanisms that govern spatial and temporal patterning during development [269]. WUS transcription is a central checkpoint in stem cell control, integrating information from disjoint regulatory pathways. Specifically, two negative feed-back loops, the WUS/CLV and the WUS/AG, seem to converge at a specific region in the WUS promoter, linking stem cell control in the SAM domain and patterning in floral organs.

The regulatory network underlying stem cell regulation and floral patterning is the target of this analysis. Computational analysis of the promoter sequences of sets of target genes, identified using experimental transcriptional profiling, reveals putative functional regulatory sequences. The sequence motif $ACATGT$ was found to be over-represented in putative WUS target genes. This motif is very similar to a TFBS of the bHLH-Zip structural class. This class contains TFs that play key roles in early cell differentiation and developmental processes [107]. The second motif $TGCATGCA$ is identical to the binding site of the CREB protein. This TF binds to the cAMP response element (CRE) and activates gene transcription in response to a wide variety of extracellular signals including growth factors and hormones. A further motif $TGGGCCT$ is found to be over-represented in WUS and LFY target genes and appears to be novel since no known matching TFBS has been reported to date. The motif is of high interest since it could be the missing link explaining the coordinated control of the stem cell population and development of leaf primordia.

The presented motifs represent a selection of the several motifs found to be over-represented in the gene lists. Typically, the motifs match known TFBSs which are implicated in processes related to growth and development. Plant TFs are poorly represented in public databases, which induces one uncertainty to the extracted knowledge. Matching plant TFBSs to vertebrate TFBSs is useful for identifying the structural TF class, however, transferring the function of the corresponding TF is not reliable since the lineage split occurred before the development of multicellularity. Although great care is taken in the experimental design and expression analysis, the gene lists are likely to include secondary responses in addition to the primary responses. Hence, it is important to acknowledge the fact that motifs may be over-represented in subsets of the gene lists rather than in all promoters. These results provide information that is needed for understanding stem cell homeostasis in plants. Further computational evaluation and experimental verification of the high confidence patterns, e.g. using reporter constructs, is needed in order to confirm these motifs and hypotheses.

6 Gene regulation at the molecular level

Sequence-based approaches facilitate analysis of the organization of regulatory regions and identification of putative TFBSs, as illustrated in the previous chapter. Statistical methods are used for representing position dependent base preferences within the binding site for a specific TF. Still, current TFBS models do not suffice for explaining observed redundancy or contradictions at the sequence level. Hence, in order to better understand the specific interactions underlying protein-DNA recognition, it is necessary to take the step from the sequence level to the molecular level [105].

Structure-based studies at atomic resolution make it possible to study the structural and chemical complementarity between TFs and their binding sites. Experimental techniques enable affinity measurements of the effects that individual mutations have on the binding free energy [183], whereas theoretical simulation approaches facilitate the analysis of such effects *in silico*. However, there are a few drawbacks associated with structure-based methods. Experimental methods that provide structural data of biomolecules are time-consuming, hence the amount of existing protein-DNA complexes with associated mutant structures is still relatively limited. The computational protocols required for analyzing a single protein-DNA complex are not trivial to design and often imply high computational requirements, which currently impede large-scale structure-based analysis. Nevertheless, the detailed knowledge gained from individual case studies of protein-DNA interaction at the molecular level is useful.

The aim is to acquire a better understanding of protein-DNA interactions, which can be used to improve the accuracy of sequence-based methods and thereby enhance both TFBS representation and discovery. An essential step on this way is to reliably estimate and predict the effects mutations made to individual bases or amino acids have on the overall binding affinity. A computational protocol, which can be used for reproducing experimental binding affinity data for ZF TFs, was developed in this study. Ideally, data obtained from large-scale simulations of protein-DNA complexes using such a protocol could be used for improving PSSM models.

The TF-DNA complex, consisting of the ZF Zif268 bound to DNA, used in this study at atomic detail, was chosen for a number of reasons. The structural class of ZF TFs is one of the most numerous in eukaryotes, members of this group have often been identified as proto-oncogenes (key regulators in cancer), extensive experimental data both sequence- and structure-based exist, and they are being explored as a framework for design of engineered TFs with putative use in therapeutics. Furthermore, the fairly small and modular recognition motif (30 amino acid residues in each finger) is well-suited for MD simulations.

6.1 Free energy calculations for Zinc Finger-DNA complexes

6.1.1 Introduction

Protein-DNA recognition is dependent on the structural and chemical complementarity at the interaction surface between the two molecules. The binding free energy of a protein-DNA complex is determined by the sum of the individual contributing factors, which either favor or oppose complex formation. Computational advances have made free energy calculations for relatively large biomolecules feasible. It is now feasible to analyze how modifications made to a chemical group at the binding interface influences the binding free energy, something which previously have been addressed using experimental approaches. A key challenge in this type of analysis is the development of computational protocols applicable to a wide range of biomolecular complexes for studying the effects of several types of modifications. If such protocols are able to reproduce experimental binding affinity data, they can be used for large-scale generation of binding free energy profiles.

A first step towards performing reliable computational calculations of binding free energies in protein-DNA complexes has been taken. The effects that small base modifications have on the overall binding affinity have been experimentally measured [40, 177]. Recently, computational estimations reproducing such effects using MD simulations and free energy calculations were reported. Sen and Nilsson [228] explored the effects of base modifications in the EcoRI-DNA complex (restriction enzyme) and Saito and Sarai [215] were able to computationally estimate experimental values for mutating thymine (T) → uracil (U) in the λ-repressor-DNA complex (TF). These studies describe a way to estimate the effects of different base modifications, however, from a transcriptional point of view it is also desirable to be able to analyze the effects of modifications made to the amino acid side chains. In this work, a computational protocol applicable to transcription factor-DNA complexes is presented.

Experimental rational design of ZF proteins has recently proven useful in targeting specific DNA sequences [216], which make them useful for a wide range of biotechnological applications [151]. Artificial ZF TFs have for example been used for inducing angiogenesis, which can accelerate wound healing [203]. Their modularity, i.e. linking several ZF domains to each other, facilitates recognition of unique DNA sequences in the genome [9, 114]. ZFs bind to their target DNA sequences without severe conformational changes and the modular binding mode of the recognition helix, which are two key features that make ZFs a suitable system for MD simulations [71]. The ZF Zif268, and mutants thereof, complexed with DNA have been subject to several experimental and theoretical studies, which address the binding affinity [40, 177]. Hence, these crystal structures are used as starting structures for the free energy calculations presented here.

Zif268 consists of three ZF domains (fingers F1, F2, and F3), where each domain contacts a four bp subsite (one bp overlap) as described in Fig. 6.1. The contacting amino acids reach out from the ZF recognition helices to make specific contacts with the bp edges in the major groove of the DNA and do also make non-specific contacts with the DNA backbone. There are several important interactions between neighboring side chains in a finger, which orient and stabilize the contacts to DNA. In each ZF domain, the zinc ion is tetrahedrally coordinated by four conserved residues, two cysteines from the β-sheet and two histidines from the α (recognition)-helix. It is primarily the residues at the N-terminus (positions -1, 2, 3, and 6 relative to the start of the helix) of each α-helix that make direct contacts with DNA. These contacts are mainly histidine-guanine and arginine-guanine interactions along one DNA strand. However, base preferences have also been observed at base positions where there are no direct interactions. A detailed analysis of the crystal structure revealed subtle contacts that allow for indirect readout of the DNA sequence [71]. Here, the acidic residues play an important role by forming water-mediated hydrogen bonds with the DNA helix. The aspartic acids at position 2 of each α-helix form hydrogen bond with the arginine at position -1. Each of these arginines (Arg18, 46, and 74) forms two hydrogen bonds with guanine. The aspartic acid-arginine contacts presumably help orienting the long side chain of arginine, increasing the arginine-guanine specificity. Furthermore, the coupled arginine/aspartic acid residue pair forms water-mediated contacts with the cytosine paired to the guanine directly contacted by the arginine residue. F1 and F3 have glutamic acid, whereas F2 has a histidine at position 3 of the recognition helix. These glutamic acids do not make any direct contacts with the DNA, however, contribute to the specificity by orienting neighboring residues and water-mediated contacts. F1 and F3 have an arginine at position 6 of the α-helix, which form hydrogen bonds with guanine. In F2, there is a threonine in position 6 that forms no direct DNA contacts, but does make water-mediated contacts with phosphate 4 of the backbone.

In this study, a computational protocol for analyzing energetic effects of base modifications at the binding interface of the TF-DNA complexes is developed. The aim is to computationally reproduce experimentally measured binding affinities. This is explored using an explicit solvent model that is implemented in the AMBER molecular modeling package [44]. Specifically, molecular modeling is used for estimating the change in the binding free energy induced

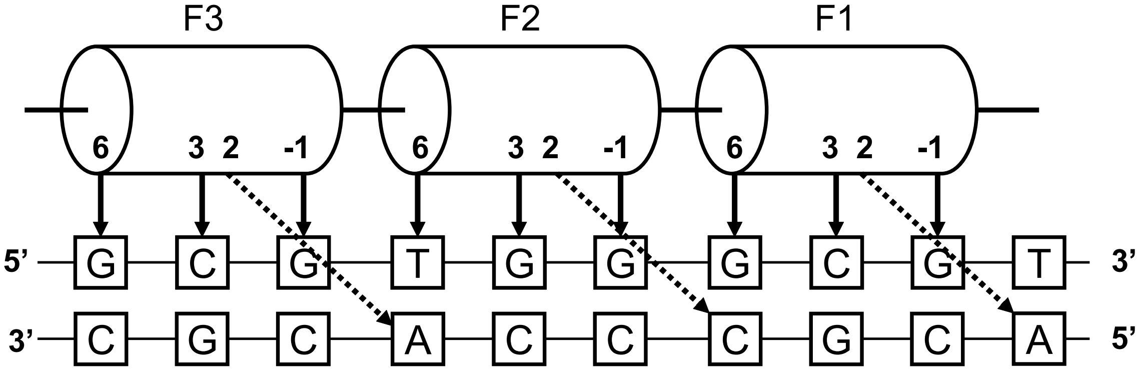


Figure 6.1: This is an illustration of the Zif268 WT protein (fingers F1, F2, and F3) binding to its consensus TFBS. The modifications were made to the third subsite, which is contacted by F1. The amino acid side chains in position -1, 2, 3, and 6 relative to the α-helix primarily contact one strand of the DNA. The side chains in F1 are Arg18, Asp20, Glu21, and Arg24, respectively.

by replacing the methyl group in the base T to a single hydrogen in base U in the WT and D20A mutant of Zif268 bound to DNA. The free energy calculations in explicit solvent are carried out allowing for full flexibility (all degrees of freedom). The TI-MD (thermodynamic integration using molecular dynamics) method and the thermodynamic cycle approach are employed for estimating the change in binding free energy (methods described in Sect. 2.3.4).

First, the computational protocol for analyzing the effects of the T → U mutation is validated using the λ-repressor-DNA complex, for which the results are compared to those reported (both experimental and computational) by Saito and Sarai [215]. The applicability of this protocol to ZF-DNA structures is further evaluated by comparing the computational results obtained here to previously reported experimental values measured by Miller *et al.* [177]. The experimental data includes measurements for mutating a T to a U within the third subsite (contacted by F1) in the complexes of the WT Zif268-DNA complex and the D20A-DNA complex. Miller *et al.* shown that the D20A mutant binds with similar affinity, but is less specific than the WT protein.

A computational protocol for mutating T → U in protein-DNA complexes using an explicit representation of the solvent molecules is described. This protocol shows similar if not better results compared to a previously presented protocol for mutating T → U in the λ-repressor-DNA complex. Applied to ZF-DNA complexes, the calculated values of the change in binding free energy agree well with previously presented experimental data. The binding free energies converge after 40 ps of sampling time and are within 0.5 kcal/mol range of the experimentally measured values. The standard deviations (STDs) show that there is some fluctuations of the estimated values, partly due to the relatively short equilibration and simulation times. The initial preparations of the structures clearly have an effect on the accuracy of the results and the required computational time. If the original crystal structure is manually mutated, longer energy minimization and equilibration times are needed.

Estimating the effects of base modifications made to protein-DNA complexes illustrates how the contributions of different components at the binding interface can be studied. Applying this protocol to novel protein-DNA complexes shows great promise for analyzing the effects on binding affinity, as it is possible to reproduce previously reported experimental values. The importance of combining experimental data and theoretical estimations at an early stage is essential for developing reliable computational protocols with applicability in rational design of artificial TFs. This study is an important first step towards enabling the use of computational methods for improving PSSM models.

6.1.2 Materials and methods

Protein-DNA complexes

Five different TF-DNA complexes are used in this study. One complex is used for comparing the performance of this protocol to a previously presented protocol and the other four complexes are used for performing computational estimations of experimentally measured binding free energy values. The effect on the binding energy due to the T → U mutation in the λ-repressor-DNA complex (position 15) has been experimentally measured to +1.80 kcal/mol [219]. Saito and Sarai [215] successfully calculated the change in binding free energy due to the mutation to +1.47 (±0.40) kcal/mol, using MD simulations in explicit solvent. The λ-repressor-DNA complex is used for validating the computational protocol developed in this study. The initial λ-repressor-DNA complex was obtained from the experimentally determined crystal structure (PDB code: 1LMB [19]). The DBD domain of the λ-repressor that binds to the consensus 12-mer of the DNA was extracted and prepared according to the instructions described by Saito and Sarai [215]. The λ-repressor-DNA complex is illustrated in Fig. 6.2, where the base T to be mutated is highlighted in red. The hydrophobic methyl group of T favorably interacts with the λ-repressor.

Miller *et al.* presented experimental results for the same modification T → U, but for a different set of protein-DNA complexes. Two different ZF proteins were used, the WT Zif268 bound to two different DNA sequences (GCG-TGG-GCT-G and GCG-TGG-GTG-G) and

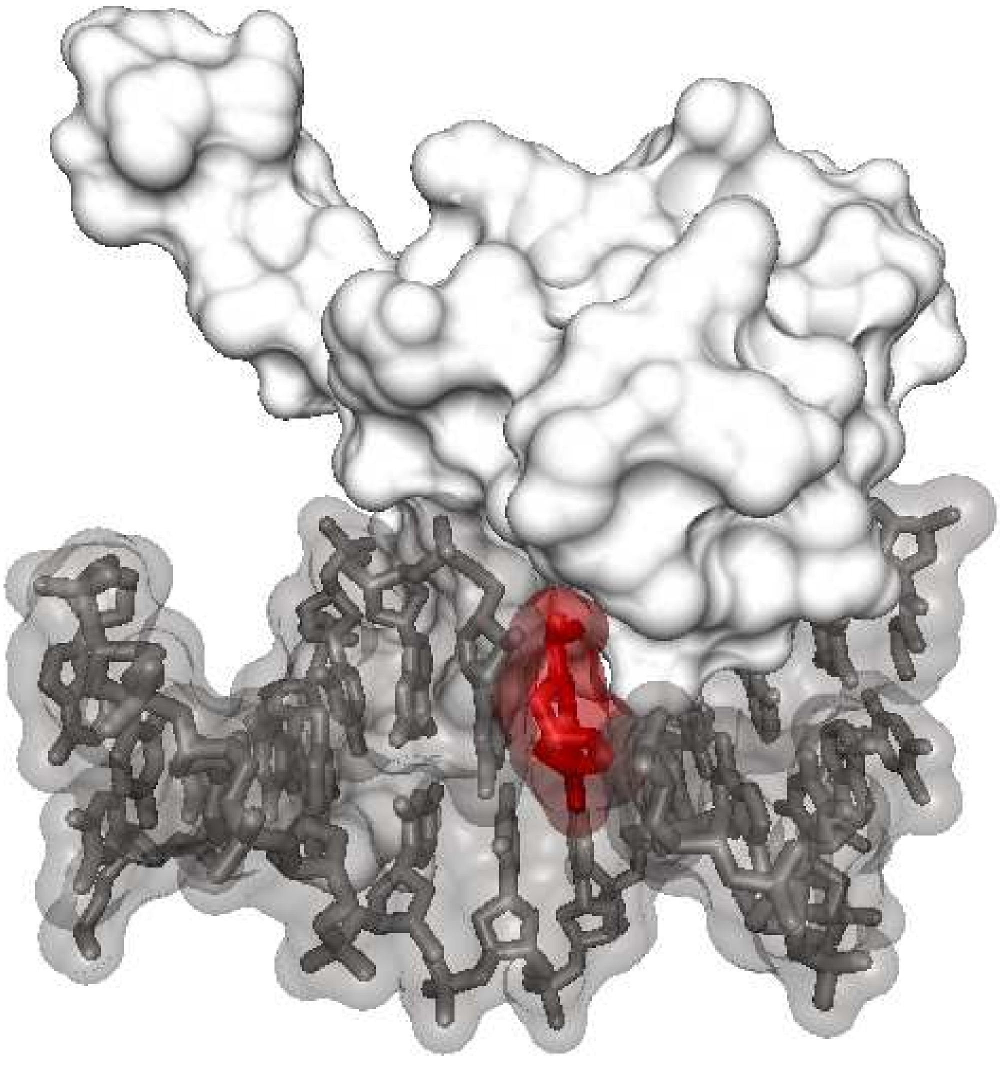


Figure 6.2: The λ-repressor-DNA (the protein is colored white and the DNA grey) complex was used by Saito and Sarai for studying the effects of the T (red) to U mutation. The mutation decreases the binding affinity, which is shown both experimentally and computationally [215].

Table 6.1: The initial modifications made to the original PDB structures in order to obtain the complexes M1, M2, M3, and M4 used by Miller *et al.* are described here. The original DNA sequence (GCG-TGG-GCG-T) in 1AAY required mutation (mutated bases are underlined and affect both DNA strands) in order to obtain the experimental structures M1 and M2. The complexes M3 and M4 required less pre-processing as they are identical, respectively very similar, to the PDB structures 1JK2 and 1JK1. Miller *et al.* experimentally measured the relative binding free energies for mutating the base T (shown in bold) to a U [177].

ID	**Protein**	**PDB code**	**DNA**
M1	WT	1AAY	GCG-TGG-GC**T**-G
M2	WT	1AAY	GCG-TGG-G**T**G-G
M3	D20A	1JK2	GCG-TGG-GC**T**-G
M4	D20A	1JK1	GCG-TGG-G**T**G-G

the D20A Zif268 mutant bound to the same two DNA sequences as the WT. These four protein-DNA complexes (M1-M4) are used for calculating the change in free energy due to replacing the methyl group in T with a hydrogen in U. The mutations are made to the T bases within the third subsite (contacted by F1, see Fig. 6.1). The four ZF-DNA complexes M1-M4 are described in Table 6.1.

Complexes M1 and M2 were obtained by mutating the WT structure bound to its consensus sequence GCG-TGG-GCG-T (PDB code: 1AAY [71]), which required two initial bp mutations. The M3 complex is the D20A mutant bound to GCG-TGG-GCT-G (identical to PDB code: 1JK2 [177]). Finally, the M4 complex was obtained from a similar complex (PDB code: 1JK1 [177]), which required one bp mutation. The initial bp modifications made to the original crystal structures in order to obtain M1-M4, are underlined in Table 6.1 and the T to be mutated is highlighted in bold. The M1 complex, with the affected T highlighted in red, is graphically illustrated in Fig. 6.3. The complexes were allowed to adjust to the structural changes, by conducting a step-wise (reducing the structural constraints) energy minimization and equilibration. Thereafter the free energy calculation is performed in order to estimate the effects of the T → U mutation.

Base mutation

Transforming T into U involves replacing the methyl group in T with a single hydrogen in U, a mutation which is graphically illustrated in Fig. 6.4. The methyl group is more hydrophobic than a single hydrogen and the mutation changes the chemical complementarity

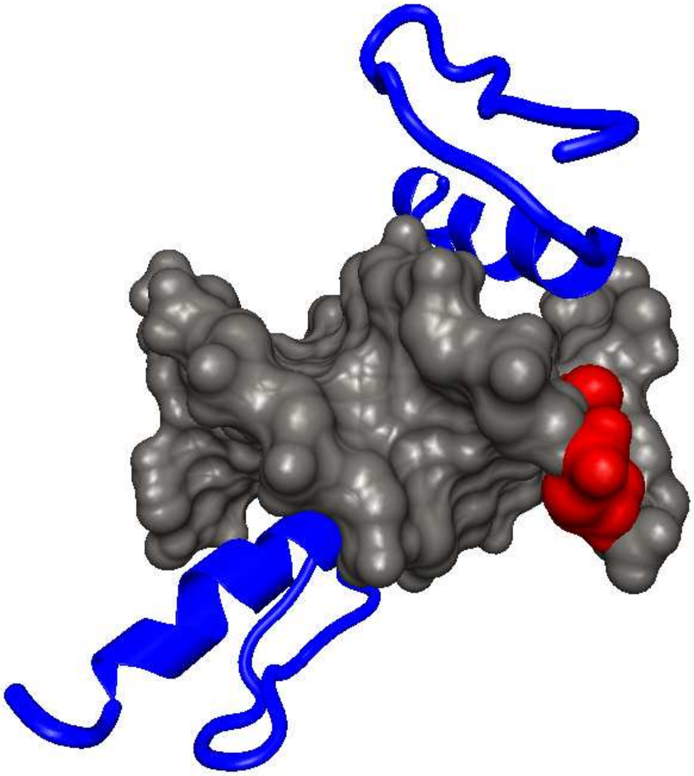


Figure 6.3: The base T (red) to be mutated to a U in the WT Zif268 (blue) DNA binding site (grey) of the M1 complex.

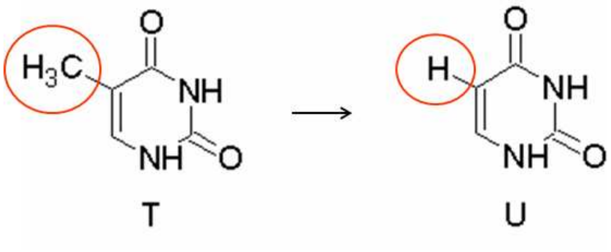


Figure 6.4: Mutating T → U involves replacing the methyl group in T group with a hydrogen in U.

at the interaction interface, which is likely to affect the binding energy. A 3D illustration of the methyl group in T and its placement within the major groove of the DNA helix is shown in Fig. 6.5.

The T → U mutation in the λ-repressor-DNA complex (Fig. 6.2) decreased the binding affinity as the $\Delta\Delta G_{Exp} = +1.80$ kcal/mol is positive. Removing the hydrophobic methyl group at the binding interface caused loss of the favored van der Waals interactions with several non-polar amino acid residues of the λ-repressor [215], especially G46, A49, and I54.

The experimental values ($\Delta\Delta G_{Exp}$) obtained for the T → U mutation in the ZF-DNA complexes by Miller *et al.*, are listed in Table 6.3. The $\Delta\Delta G_{Exp}$ for the complexes M1, M2, M3, and M4 are -0.09, -0.17, +0.08, and -0.76 kcal/mol, respectively. These values are relatively low compared to the +1.80 kcal/mol obtained for the λ-repressor-DNA complex. The T in the λ-repressor-DNA complex is at a crucial position at the binding interface, hence affects the binding free energy more.

Free energy calculations and protocols

The effects of the base mutations were estimated using an explicit solvent model (described in Sect. 2.3.3). The AMBER 7.0 molecular modeling package [44] with the AMBER99 force field [266] is used throughout this study, since these have been shown to be useful for molecular modeling studies involving proteins and DNA.

Explicit solvent

Computationally mutating T → U using explicit solvent involves defining a hybrid base (TU). The hybrid contains structural information about both T and U and a mixing parameter λ allows for a smooth transition between the two states, starting as the base T (*unperturbed state*) and finishing as U (*perturbed state*). This information is stored in parameter and

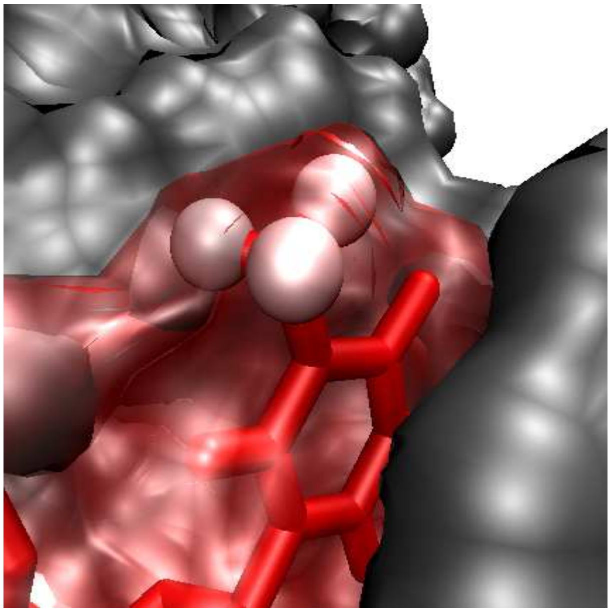


Figure 6.5: The hydrogens (white) of the affected methyl group of the base T (red) in the major groove of the DNA binding site (grey).

library files, that were specifically constructed for this purpose. The change in the binding free energy (ΔG) is calculated using the TI-MD method of the AMBER package, which is a commonly employed technique for computing the change in binding free energy due to a structural modification. The thermodynamic cycle approach, which was described in Sect. 2.3.4 and illustrated in Fig. 2.13, gives the following relationship for obtaining the relative change in binding free energy:

$$\Delta\Delta G = \Delta G_2 \text{ - } \Delta G_1 = \Delta G'' \text{ - } \Delta G'$$

For each protein-DNA complex two parallel transformations, going from the unperturbed ($\lambda = 0$) to the perturbed ($\lambda = 1$) state, one for calculating $\Delta G''$ (*associated state*) and $\Delta G'$ (*dissociated state*) transformations, are performed. Each of these are conducted by simulating the complex at several intermediate states (integration points) defined by λ according to the following approximation:

$$\Delta G \approx \sum_{i=1}^{n} w_i \langle \delta V / \delta \lambda \rangle_{\lambda_i}$$

The ensemble average change in potential energy V $\langle \delta V / \delta \lambda \rangle$ between two intermediate steps of λ (λ_i andλ_{i+1}) is estimated using MD simulations. At each integration point (λ_i) the $\langle \delta V / \delta \lambda \rangle$ value is multiplied by a weight w_i that is symmetrical around $\lambda = 0.5$. Quadrature points (i) are used for estimating the integral and the weights w_i are defined according to the trapezoidal rule (the curious reader is referred to [2, 44] for further details).

According to the definition, two TI-MD simulations ($\Delta G'$ and $\Delta G''$) are required for estimating the $\Delta\Delta G$. For each complex (M1, M2, M3, and M4) the associated state was prepared by adding explicit hydrogens to the crystal structure. The protonation state of ionizable groups was determined corresponding to pH 7, however, exceptions were made for the metal ion-coordinating residues. The complex was then completely embedded in a water box, with at least 8 Å from the complex to the borders. Counterions were added in order to neutralize the system and water molecules interfering with the added counterions were removed. Since the protein is not affected by the modification, the dissociated state was prepared in a similar fashion, however, immersing only the DNA molecule in a 8 Å thick water layer.

Before starting to estimate the change in free energy due to a modification it is important that the system reaches thermodynamic equilibrium. Especially artificial mutations made to the starting systems are likely to introduce unfavorable contacts. Excessively high energies typically originate in van der Waals or electrostatic interactions between non-bonded atoms that have come to close. Performing two essential pre-processing steps: *minimization* and *equilibration* prevents high initial potential energy from being transformed into high kinetic energy. An energy minimization seeks to relieve any energetically unfavorable clashes in the system, by bringing the system to a nearby local energy minimum. The minimization is performed in two steps. First, 500 steps of minimization restraining heavy atoms, including 50 steps of steepest descent (SD) followed by 450 steps of conjugate gradient (CG), were performed. The heavy atoms are restrained in order to let the energy of the explicitly added water molecules around the solute molecule decrease. Secondly, a non-restrained minimization was carried out during a total of 1000 steps, which included 50 SD and 950 CG steps. The energy decreases quickly in the beginning and then levels off at the end of the minimization, as the system reaches the local minimum. The aim of the equilibration is to bring the system to a physically meaningful conformation by allowing sampling of possible conformations and achieve thermodynamic equilibrium. Equilibration was carried out in two steps. First, the system is coupled to a heat bath of 300 K (raising the temperature from 0 K to 300 K) and equilibrated for 5 ps under constant volume (NTV) dynamics, aiming to let the temperature stabilize at the desired value. In the second equilibration step the temperature is kept at 300 K and it is carried out under constant pressure (NTP) dynamics for another 5 ps. A weak harmonic constraint of 1 kcal/mol was applied to the heavy atoms in both steps, whereas the surrounding solvent molecules are allowed to adjust to the solute. The non-bonded interactions are evaluated up to the cutoff (cut) set to 10 Å. SHAKE constraints are used on the hydrogen atoms ($ntc = 2, ntf = 2$).

The TI simulations are run in a step-wise fashion, going from the start state (base T) to the end state (base U), corresponding to $\lambda = 0$ and $\lambda = 1$, respectively. Two different experimental protocols are tested. The first (Explicit 1, E1) includes five integration points (λ_i), whereas the second (Explicit 2, E2) includes nine. The protocol E1 includes 20 ps of equilibration time (no artificial restraints were applied) before 50 ps of simulation time, where the sampling of $\langle \delta V/\delta\lambda \rangle$ is conducted, at each integration point. In the second protocol (E2) the equilibration time is increased to 50 ps.

6.1.3 Results

Validation

The λ-repressor-DNA complex is used for validating the computational protocols used for performing the T → U mutation. The experimentally measured change in the binding free energy was +1.80 kcal/mol and computationally estimated by Saito and Sarai to +1.47 (±0.40) kcal/mol, which is listed in Table 6.2.

Protocol E1 included 20 ps equilibration and 50 ps simulation time at five integration points. In protocol E2 the equilibration time was increased to 50 ps and the number of integration points was increased to nine. The results obtained using E1 and E2 generally show good agreement with experimental data and previously reported values, see in Table 6.2. In particular, the $\Delta\Delta G_{E2}$ value of +1.83 (±1.90) kcal/mol is very close to the experimental value, accurately estimating the loss of favored van der Waals interactions. The STDs are analyzed two-fold, first by including data from 50 ps of the simulation time at each integration point, and secondly by using only the final 10 ps at each integration point. In the first case the STDs are rather high at a value of ±1.90 kcal/mol, however, during the final 10 ps the complex has better adjusted to the new conformation and the STD is ±0.77 kcal/mol, as listed in Table 6.4. The STDs are within the range of previously reported results [228].

Comparison against experimental data

In this second test the protocols are used for estimating the effect of the T → U mutation for the four ZF-DNA complexes (M1-M4), for which experimental results but no computational estimations have been reported. The results obtained when applying the above described

Table 6.2: The results of the performance evaluation of the computational protocols (E1 and E2) using explicit solvent are listed in this table. The values are listed next to the experimentally measured and previously reported computational result (Experimental and Saito, respectively). E1 includes 20 ps equilibration and 50 ps simulation at five integration points, whereas E2 includes 50 ps equilibration and 50 ps simulation at nine integration points. Values are obtained from 50 ps of simulation time ([0..50]).

Protocol	$\Delta\Delta G$	ΔG	**Complex**
	(kcal/mol)	(kcal/mol)	**λ-repressor-DNA**
Exp.	$\Delta\Delta G_{Exp}$		+1.80
Saito	$\Delta\Delta G_{Saito}$		+1.47 (±0.40)
Explicit 1	$\Delta\Delta G_{E1}$		+1.00 (±2.14)
[0..50]		$\Delta G_{G''}$	-5.77 (±0.90)
		$\Delta G_{G'}$	-6.77 (±1.24)
Explicit 2	$\Delta\Delta G_{E2}$		+1.83 (±1.90)
[0..50]		$\Delta G_{G''}$	-4.82 (±0.99)
		$\Delta G_{G'}$	-6.64 (±1.14)

protocols (E1 and E2) and evaluating over the full simulation time are listed in Table 6.3. Again, the results obtained using the E2 protocol are closer to the experimentally reported values. The relative free energies for the complexes M1-M4 are plotted against simulation time in Fig. 6.6. The results converge for longer simulation times after about 40 ps. The results are listed two-fold: First, the two protocols E1 and E2 are evaluated over the full 50 ps of simulation time and the results are listed in Table 6.3. Secondly, the results evaluated using only the final 10 ps of simulation time are listed in Table 6.4.

Especially, the $\Delta\Delta G_{E2ii}$ values for the complexes M3 and M4 (-0.17 (±0.75) kcal/mol and -1.02 (±0.92) kcal/mol) are near to identical to the experimental values (-0.08 (±0.14) kcal/mol and -0.76 (±0.81) kcal/mol, respectively) (see Table 6.4). The computational results for M1 also agree with the experimental values, which are -0.09 (±0.13) kcal/mol and -0.68 (±0.71), respectively. The results for M2 do not agree as the experimental result is -0.17 (±0.25) kcal/mol and the computational +2.10 (±0.62). Longer equilibration and simulation times are expected to slightly improve the results, especially for the complexes M1 and M2 that needed larger initial modifications. The results for the complexes M1 and M2 are listed in Table 6.3 and in Table 6.4. The STDs during the final 10 ps of simulation time is between ±0.62 and ±0.92 kcal/mol, which is within the range presented by Sen and Nilsson [228].

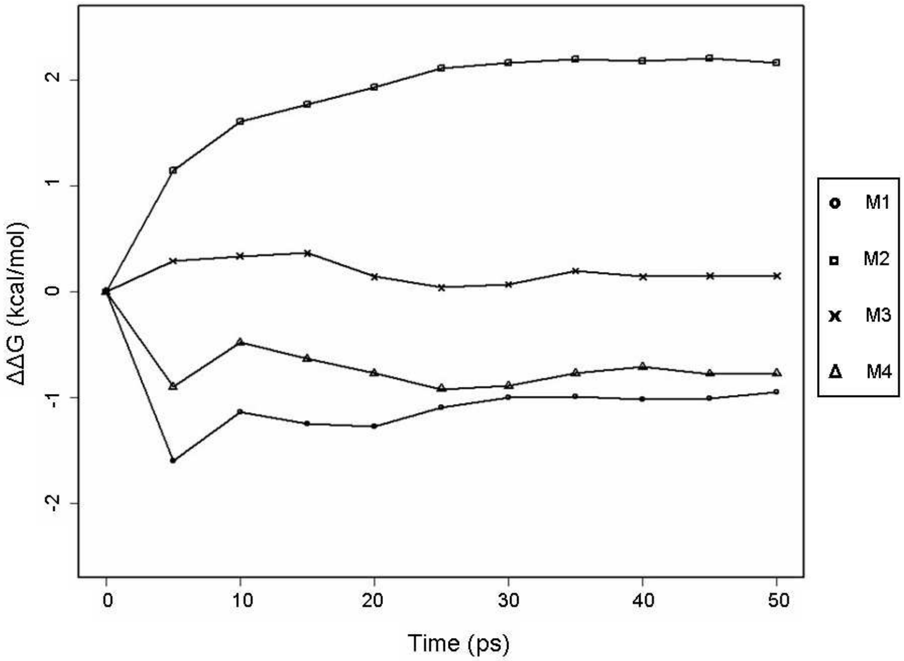


Figure 6.6: Free relative free energy for each complex M1-M4 is plotted against simulation time. Convergence is achieved at about 40 ps of simulation time. Slight fluctuations indicate that longer simulation times could improve the results even further.

Table 6.3: The computationally calculated results for the T → U mutation, using the protocols E1 and E2 for the structures M1-M4, are compared to the respective experimental values. The STDs are calculated for the full 50 ps of simulation time and shown in parenthesis.

Protocol	$\Delta\Delta G$	ΔG	**Complex**							
	(kcal/mol)	(kcal/mol)	**M1**		**M2**		**M3**		**M4**	
Exp.	$\Delta\Delta G_{Exp}$		-0.09 (±0.13)		-0.17 (±0.25)		+0.08 (±0.14)		-0.76 (±0.81)	
Explicit 1	$\Delta\Delta G_{E1}$		-2.19 (±2.54)		+1.96 (±1.73)		+0.77 (±2.02)		+1.06 (±1.88)	
[0..50]		$\Delta G_{G''}$		-8.67 (±1.87)		-4.73 (±1.02)		-5.81 (±1.50)		-5.61 (±1.11)
		$\Delta G_{G'}$		-6.47 (±1.47)		-6.67 (±1.43)		-6.50 (±1.41)		-6.72 (±1.52)
Explicit 2	$\Delta\Delta G_{E2}$		-0.95 (±2.09)		+2.17 (±1.89)		+0.15 (±2.10)		-0.77 (±2.11)	
[0..50]		$\Delta G_{G''}$		-7.39 (±1.57)		-4.25 (±1.12)		-6.53 (±1.54)		-7.86 (±1.35)
		$\Delta G_{G'}$		-6.44 (±1.40)		-6.41 (±1.37)		-6.68 (±1.57)		-7.08 (±1.56)

Table 6.4: The relative free energies converge after 40 ps and are listed here. The corresponding standard deviations are significantly smaller, when evaluated for the final 10 ps of the simulation time.

Protocol	$\Delta\Delta G$			**Complex**		
	(kcal/mol)	**λ-repressor-DNA**	**M1**	**M2**	**M3**	**M4**
Exp.	$\Delta\Delta G_{Exp}$	+1.80 (±na)	-0.09 (±0.13)	-0.17 (±0.25)	+0.08 (±0.14)	-0.76 (±0.81)
Explicit 2i [0..50]	$\Delta\Delta G_{E2i}$	+1.83 (±1.90)	-0.95 (±2.09)	+2.17 (±1.89)	+0.15 (±2.10)	-0.77 (±2.11)
Explicit 2ii [40..50]	$\Delta\Delta G_{E2ii}$	+2.00 (±0.77)	-0.68 (±0.71)	+2.10 (±0.62)	+0.17 (±0.75)	-1.02 (±0.92)

6.1.4 Discussion

Crystal structures provide snapshots of macromolecular interactions and can help to understand the thermodynamics underlying both specific and non-specific interactions. Complex formation should be seen as a concerted event that is dependent on several factors. Thermodynamic features involved in protein-DNA binding include electrostatic interactions, shape complementarity, and solvent release, which are all considered to be favorable to the binding process. On the other hand, entropy effects: the loss of translational, rotational, and internal degrees of freedom due to complex formation, are generally unfavorable. Computational methods serve as a means for analyzing the implications structural modifications have on the binding free energy. However, should be seen in the light of their limitations and as a complement of (rather than as a replacement) to current experimental methods.

In this study, the challenge of analyzing the effects structural modifications have on the thermodynamics of binding has been addressed using the TI-MD approach. The explicit solvent protocol presented here, considers the relative effects of the modifications between two states, the associated and the dissociated, by performing free energy calculations and by applying the thermodynamic cycle approach. The free energy calculations were performed under all degrees of freedom including explicit modeling of water molecules, addressing some important issues with presently available theoretical methods.

In the comparison against an existing protocol an improved performance is shown for the λ-repressor-DNA complex. The developed protocol, applied to the ZF-DNA complexes, is proven useful for estimating the change in binding free energy due to the T → U mutation. In the λ-repressor-DNA complex the thymine methyl group is at the center of the binding interface and interacts with two non-polar amino acids A49 and I54. Hence, the mutation causes loss of favorable hydrophobic interactions (positive $\Delta\Delta G$).

Results that are very close to the experimentally reported ones are obtained when performing the mutation in the D20A mutant-DNA complexes (M3 and M4). In complex M3, the thymine group interacts with the hydrophobic side chain of A20. Removing the methyl group of thymine in M3 leads to a loss of favorable hydrophobic interactions this is reflected in the computational result $\Delta\Delta G_{E2ii} = +0.17$ (±0.75) kcal/mol, which is further strengthened by the experimental result $\Delta\Delta G_{Exp} = +0.08$ (±0.14) kcal/mol.

In the M4 complex the closest amino acid is the G21 side chain, which steps in and orients the R18 side chain in absence of D20 (however, less constrained). Hence, mutating T → U improves the binding affinity. The computational value $\Delta\Delta G_{E2ii}$ = -1.02 (±0.92) kcal/mol is again in agreement with the experimental value $\Delta\Delta G_{Exp}$ = -0.76 (±0.81) kcal/mol. In M1, the thymine methyl group interacts with D20. Replacing the methyl group with an hydrogen improves the binding affinity. The computational value $\Delta\Delta G_{E2ii}$ = -0.68 (±0.71) kcal/mol suggests that it indeed is a favorable mutation, however, it is an slight over-estimation of the experimentally measured effect $\Delta\Delta G_{Exp}$ = -0.09 (±0.13) kcal/mol. The T → U mutation in the M2 complex was experimentally measured to -0.17 kcal/mol, however, for this complex the computational protocol fails to estimate the negative effect on the binding free energy. It is likely that the relatively large initial mutations made in order to obtain the starting complex (described in Table 6.1) complicate a correct estimation of the binding free energy. The calculated STDs are still slightly higher than those obtained from experimental measurements, but in the range of those previously reported in similar studies [215, 228].

The results confirm earlier suggestions that it is possible to estimate changes in the binding free energies in protein-DNA complexes by performing MD simulations. Furthermore, they indicate that the accuracy is highly dependent on the initial complex. The larger the initial modifications needed in order to obtain the starting structure, the less reliable results are to be expected. The largest structural modifications were done in order to construct the starting complexes M1 and in particular M2, for which the calculated results deviate the most from the experimentally obtained ones. This can most likely be compensated by extending the equilibration and simulation times, which has been called for by others [91].

Choosing the solvent model to use is clearly important and analysis specific problem. Implicit solvent models approximate the interactions with the bulk solvent, thereby reducing the computational costs. Such models are likely to give more reliable results than explicit solvent models, when either the initial modifications in order to obtain the starting complex, or the structural mutation itself is larger. As these calculations are faster, they are also better suited for efficient screening of larger data sets. A study, in which an implicit solvent model is applied to the same data set by Miller *et al.*, is ongoing. Preliminary test show that implicit solvent models do not handle the DNA molecule as well as explicit models, since DNA strand separation has been observed. However, by applying simulation constraints these problems can be circumvented. Implicit models have one great advantage, as iterations of full bp mutations are feasible for one protein-DNA complex. The aim is to compare the two methods and bring the results together in order to evaluate required simulation times and applicability for improving PSSM models.

The computational protocol presented here is suited for analyzing modifications made to protein-DNA complexes. The contributions of individual residues at the binding interface can be studied, whereby residues responsible for the key interactions can be identified. Results that agree well with experimental values can be obtained at reasonable simulation times if the mutation to be studied is small and the initial structure is close to a crystal structure. Analysis of a larger data set is likely to provide useful information and a rough estimation of the effects, however, currently hindered by the high computational requirement using explicit solvent representation and missing experimental data for validation.

The aim is to contribute to the development of predictive methods of protein-DNA interactions by including full flexibility, different solvent models, and atomic description of the complexes. In order to achieve this goal, the first step towards developing computational algorithms for accurate prediction of TFBSs is to understand the recognition process qualitatively. Theoretical approaches for performing virtual mutagenesis using thermodynamic cycles provide insight into macromolecular energetics, which is essential step in rational design of sequence specific ZF proteins and for investigating potential use in therapy.

7 Conclusion

Gene expression is primarily regulated at the level of transcription through sequence specific recognition of TFBSs. Intertwined signaling pathways and gene regulatory networks care for controlled global expression profiles of genes in all cell types. Due to their role in gene regulation, TFs play key roles in cellular events, such as cell survival, growth, and differentiation. Consequently, alterations in genes encoding TFs are implicated in a number of diseases.

The analysis of gene regulation at the level of systems (presented in Chapt. 4) is a good example of how multiple sources of information can be used for discovering potential causes of a complex disease. The importance of gene regulation in cancer has been brought out by providing evidence for a correlation between differential gene expression and auto immune response. Furthermore, sets of biologically related cancer specific target genes are identified using this global approach.

The three studies presented in Chapt. 5 provide instructive information and illustrate the power of sequence-based tools for predicting functional TF binding sites, which in turn can provide clues about the underlying regulatory networks. The results include a evidence for potential TFs involved in regulating the expression of melanoma-associated genes, which is confirmed by previously presented experimental results. The integrated system YRSA exemplifies the advantage of flexible and automated analysis of biologically-related genes in yeast and has proved useful for biologists active in the field. Finally, potentially functional TFBSs involved in directing stem cell regulation in plant are presented. An over-represented motif, which could be the link between the two key TFs LFY and WUS, is suggested. The genomes analyzed here differ in both size and complexity. Specifically the promoters in *H. sapiens* are significantly longer and more complex than those of *S. cerevisiae.* Phylogenetic footprinting was employed in the analysis of the melanoma-related genes, as a means for filtering out non-conserved regions, whereas the promoter sequences of both *S. cerevisiae* and *A. thaliana* can be analyzed without pre-processing. Understanding regulatory mechanisms in less complex organisms is a step towards studying the corresponding mechanisms in human.

These examples illustrate the value of studying gene regulation in model organisms.

Structure-based methods provide means for explaining contradictions that are observed at the sequence level. Significant progress in structure-based analysis of protein-DNA complexes has been made both experimentally and computationally. An initial step towards analyzing the individual contributions to the binding affinity between protein and DNA is taken in the structure-based study presented in Chapt. 6. An investigation at the molecular level facilitates a deeper understanding of the energetic effects, which are due to subtle differences in chemical and structural complementarity between the two interacting molecules. Eventually, such subtle differences determine the binding affinity and thereby the subsequent gene expression levels in the cells. In this study, the power of using MD simulations for investigating the effects that structural modifications have on the relative free energies in individual protein-DNA complexes has been demonstrated. This study is an important step towards using molecular modeling techniques for constructing PSSMs.

Gene regulation has been intensely investigated during recent years, which has resulted in a number of valuable insights into transcriptional involvement in the development and progression of diseases. Clearly information obtained at the level of systems is crucial in the drug target identification process and for diagnostic purposes. Sequence-based analyses of protein-DNA binding have contributed to the initial and necessary understanding of the biomolecular interactions involved in specific sequence recognition. Sequence-based methods are currently irreplaceable regarding their use for efficiently scanning genomic sequences. However, structure-based information is crucial for understanding specific TF-DNA binding, for improving sequence-based representation and discovery of TFBSs, and for eventually untangling gene regulation.

Technologies capable of modulating gene expression are powerful tools for biotechnological research and can be applied in therapeutics. Using engineered TFs for controlling gene expression presents some advantages compared to other tools such as antisense RNA or RNA interference (RNAi). The ZF-TF technology is currently the only method, by which transcription can be either activated or repressed. Another clear advantage is the synergistic binding effects observed when constructing a multidomain ZF protein. Finally, transcriptional regulation is present in diverse organisms and can be used in prokaryotes, as well as in higher eukaryotes. Before being able to utilize TFs in therapy, it is of outmost importance to ensure that the individual effects of a single TF and its involvement in regulatory networks can be assessed. Rationally designed artificial ZF-TFs are of potential use in therapeutics, for example in cancer therapy and for controlling the differentiation of transplanted stem cells. Members of the Cys_2His_2 ZF family bind to DNA using modular domains that are capable of recognizing a wide range of DNA sequences. Ideally, any given DNA sequence could be targeted using a ZF library consisting of sequence specific modules that can be combined to design multidomain ZF proteins. A universal recognition code specifying amino acid-base preferences is desired.

In conclusion, with the vision is to fully unravel complex regulatory networks underlying developmental decisions it is meaningful to study several aspects of gene regulation. The versatile analysis presented in this work is a good example of how gene regulation can be addressed using different approaches, which is a necessity for exploring the use of gene regulation in therapeutics. Approaching gene regulation at the level of systems provides a holistic view of the involvement of gene regulation is disease. Zooming in to the level of sequences, it is possible to identify common regulatory mechanisms. The final step down to molecular level provides the necessary information for understanding sequence specific recognition in protein-DNA complexes. Exactly how detailed structure-based information obtained at the molecular level should be used for improving sequence-based methods remains an open question. A further non-trivial problem, but perhaps the most interesting aspect of gene regulation, is how gene regulatory networks can be simulated at the level of systems using sequence-based information.

They are ill discoverers that think there is no land, when they can see nothing but sea.

Sir Francis Bacon

8 Abbreviations

bp base pair

bZip basic leucine zipper

DBD DNA-binding domain

DNA deoxyribonucleic acid

ChIP chromatin immunoprecipitation assay

GB Generalized Born model

HD homeo domain

HLH helix-loop-helix

HTH helix-turn-helix

kb kilobases

MD molecular dynamics

MM molecular mechanics

Myrs Million years

ORF open reading frame

PSSM position specific scoring matrix

RE response element

SASA solvent accessible surface area

STD standard deviation

TF transcription factor

TFBS transcription factor binding site

TI thermodynamic integration

TIC transcription initiation complex

TSS transcription start site

WT	wild type
UAS	upstream activating sequence
ZF	zinc finger
3D	three-dimensional
Å	Ångström

9 Appendix

Curriculum Vitae

Personal

Full name:	Annette Terese Höglund
Born:	Oct. 3rd, 1977, Åland Islands, Finland
Nationality:	Finnish
Languages:	Swedish, English, German, and some Finnish

Education

2002 - 2006	PhD Student, Tübingen University/Saarland University, Germany
2000	Student, MSc Project, Stockholm Bioinformatics Center, Sweden
1999 - 2000	ERASMUS Student, University of Salford, UK
1996 - 2000	Student, MSc Engineering Biology, University of Linköping, Sweden

Relevant Work Experience

2002 - 2006	Teaching experience: MSc and BSc projects, seminars, and practicals Tübingen University, Germany
2001 - 2002	Guest Lecturer, University of Linköping, Sweden
2001	Project Assistant, Center for Genomics and Bioinformatics (CGB) Karolinska Institute, Sweden
2001	Internship, Informatics Research, Celera Genomics, Rockville, MD, USA
1999	Internship, NOKIA Mutlimedia Terminals, Sweden, Germany, and Finland

List of Publications

2006 Höglund, A., Blum, T., Brady, S., Dönnes, P., Miguel, J.S., Rocheford, M., Kohlbacher, O., and Shatkay, H., Significantly improved prediction of subcellular localization by integrating text and protein sequence data. In: *Pacific Symposium on Biocomputing (PSB 2006), p. 16-27.*

Höglund, A., Dönnes, P., Blum, T., Adolph, H.-W., and Kohlbacher, O. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs, and amino acid composition. *Bioinformatics (in press).*

2005 Höglund, A., Dönnes, P., Adolph, H.-W., and Kohlbacher, O. From prediction of subcellular localization to functional classification: Discrimination of DNA-packing and other nuclear proteins. *Online Journal of Bioinformatics 6(1), p. 51-64.*

Höglund, A., Dönnes, P., Blum, T., Adolph, H.-W., and Kohlbacher, O. Using N-terminal targeting sequences, amino acid composition, and sequence motifs for predicting protein subcellular localization. In: *Proceedings of the German Conference on Bioinformatics (GCB 2005), p. 45-59.*

2004 Dönnes, P., Höglund, A., Sturm, M., Comtesse, N., Backes, C., Meese, E., Kohlbacher, O., and Lenhof, H.-P. Integrative analysis of cancer-related data using CAP. *FASEB Journal 18(12), p. 1465-1467.*

Dönnes, P., and Höglund, A. Predicting protein subcellular localization: past, present, and future. *Genomics Proteomics Bioinformatics 2(4), p. 209-215.*

Höglund, A., and Kohlbacher, O. From sequence to structure and back again: approaches for predicting protein-DNA binding. *Proteome Sci 2(1), p. 3.*

2003 Sandelin, A., Höglund, A., Lenhard, B., and Wasserman, W. Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes. *Funct Integr Genomics 3(3), p. 125-134.*

Bibliography

[1] National Cancer Institute, http://www.cancer.gov/.

[2] Abramovitz, M. *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York, US, 1974.

[3] Adam, J. K., Odhav, B., and Bhoola, K. Immune responses in cancer. *Pharmacol Ther 99*, 1 (Jul 2003), 113–132.

[4] Ajay, and Murcko, M. Computational methods to predict binding free energy in ligand-receptor complexes. *J Med Chem 38*, 26 (1995), 4953–4967.

[5] Alberts, B. *Molecular Biology of the Cell, 4th Edition*. Garland, NY London, 2002.

[6] Angelopoulou, K., Yu, H., Bharaj, B., Giai, M., and Diamandis, E. P. p53 gene mutation, tumor p53 protein overexpression, and serum p53 autoantibody generation in patients with breast cancer. *Clin Biochem 33*, 1 (Feb 2000), 53–62.

[7] Ashcroft, N. W., and Mermin, D. *Solid State Physics*. Saunders College, Philadelphia, US, 1999.

[8] Asturias, F. J. RNA polymerase II structure, and organization of the preinitiation complex. *Curr Opin Struct Biol 14*, 2 (Apr 2004), 121–129.

[9] Bae, K.-H., Kwon, Y. D., Shin, H.-C., Hwang, M.-S., Ryu, E.-H., Park, K.-S., Yang, H.-Y., Lee, D.-K., Lee, Y., Park, J., Kwon, H. S., Kim, H.-W., Yeh, B.-I., Lee, H.-W., Sohn, S. H., Yoon, J., Seol, W., and Kim, J.-S. Human zinc fingers as building blocks in the construction of artificial transcription factors. *Nat Biotechnol 21*, 3 (Mar 2003), 275–280.

[10] Bailey, T. L., and Gribskov, M. The megaprior heuristic for discovering protein sequence patterns. *Proc Int Conf Intell Syst Mol Biol 4* (1996), 15–24.

[11] Bairoch, A., and Apweiler, R. The SWISS-PROT protein sequence database and its supplement in TrEMBL in 2000. *Nucleic Acids Res. 28*, 1 (2000), 45–48.

[12] Bar-Eli, M. Role of AP-2 in tumor growth and metastasis of human melanoma. *Cancer Metastasis Rev 18*, 3 (1999), 377–385.

[13] Bar-Eli, M. Gene regulation in melanoma progression by the AP-2 transcription factor. *Pigment Cell Res 14*, 2 (Apr 2001), 78–85.

[14] Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. Modeling dependencies in protein-DNA binding sites. In *RECOMB* (2003), pp. 28–37.

[15] Barton, M. K., and Poethig, R. S. Formation of the shoot apical meristem in *Arabidopsis thaliana*: An analysis of development in the wild type and in the shoot meristemless mutant. *Development 119* (1993), 823–831.

[16] Bartsevich, V. V., Miller, J. C., Case, C. C., and Pabo, C. O. Engineered zinc finger proteins for controlling stem cell fate. *Stem Cells 21*, 6 (2003), 632–637.

[17] Baurle, I., and Laux, T. Regulation of WUSCHEL Transcription in the Stem Cell Niche of the *Arabidopsis* Shoot Meristem. *Plant Cell* (Jun 2005).

[18] Bayani, J., Brenton, J. D., Macgregor, P. F., Beheshti, B., Albert, M., Nallainathan, D., Karaskova, J., Rosen, B., Murphy, J., Laframboise, S., Zanke, B., and Squire, J. A. Parallel analysis of sporadic primary ovarian carcinomas by spectral karyotyping, comparative genomic hybridization, and expression microarrays. *Cancer Res 62*, 12 (Jun 2002), 3466–3476.

[19] Beamer, L. J., and Pabo, C. O. Refined 1.8 Å crystal structure of the lambda repressor-operator complex. *J Mol Biol 227*, 1 (Sep 1992), 177–196.

[20] Beerli, R. R., Schopfer, U., Dreier, B., and Barbas, C. F. Chemically regulated zinc finger transcription factors. *J Biol Chem 275*, 42 (Oct 2000), 32617–32627.

[21] Bell, C. E., and Lewis, M. Crystallographic analysis of Lac repressor bound to natural operator O1. *J Mol Biol 312*, 5 (Oct 2001), 921–926.

[22] Benos, P. V., Bulyk, M. L., and Stormo, G. D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res 30*, 20 (2002), 4442–4451.

[23] Benos, P. V., Lapedes, A. S., Fields, D. S., and Stormo, G. D. SAMIE: statistical algorithm for modeling interaction energies. *Pac Symp Biocomput* (2001), 115–126.

[24] Benos, P. V., Lapedes, A. S., and Stormo, G. D. Is there a code for protein-DNA recognition? Probab(ilistical)ly... *Bioessays 24*, 5 (2002), 466–475.

[25] Benos, P. V., Lapedes, A. S., and Stormo, G. D. Probabilistic code for DNA recognition by proteins of the EGR family. *J Mol Biol 323*, 4 (2002), 701–727.

[26] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. The Protein Data Bank. *Nucleic Acids Res 28*, 1 (2000), 235–242.

[27] Blackwood, E. M., and Kadonaga, J. T. Going the distance: a current view of enhancer action. *Science 281*, 5373 (Jul 1998), 60–63.

[28] Blanchette, M., and Tompa, M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res 12*, 5 (2002), 739–748.

[29] Blazquez, M. A., Soowal, L. N., Lee, I., and Weigel, D. LEAFY expression and flower initiation in *Arabidopsis*. *Development 124*, 19 (Oct 1997), 3835–3844.

[30] Bonifer, C. Developmental regulation of eukaryotic gene loci: which cis-regulatory information is required? *Trends Genet 16*, 7 (Jul 2000), 310–315.

[31] Brand, U., Fletcher, J. C., Hobe, M., Meyerowitz, E. M., and Simon, R. Dependence of stem cell fate in *Arabidopsis* on a feedback loop regulated by CLV3 activity. *Science 289*, 5479 (Jul 2000), 617–619.

[32] Brand, U., Grunewald, M., Hobe, M., and Simon, R. Regulation of CLV3 expression by two homeobox genes in *Arabidopsis*. *Plant Physiol 129*, 2 (Jun 2002), 565–575.

[33] Bratke, J., Kietzmann, T., and Jungermann, K. Identification of an oxygen-responsive element in the 5'-flanking sequence of the rat cytosolic phosphoenolpyruvate carboxykinase-1 gene, modulating its glucagon-dependent activation. *Biochem J 339* (May 1999), 563–569.

[34] Brazma, A., Jonassen, I., Eidhammer, I., and Gilbert, D. Approaches to the automatic discovery of patterns in biosequences. *J Comput Biol 5*, 2 (1998), 279–305.

[35] Breyne, P., van Montagu, M., Depicker, N., and Gheysen, G. Characterization of a plant scaffold attachment region in a DNA fragment that normalizes transgene expression in tobacco. *Plant Cell 4*, 4 (Apr 1992), 463–471.

[36] Brivanlou, A. H., and Darnell, J. E. J. Signal transduction and the control of gene expression. *Science 295*, 5556 (Feb 2002), 813–818.

[37] Brown, T. A. *Genomes, 2nd Edition*. Bios Scientific Publishers Ltd, Oxford, 2002.

[38] Buckingham, S. Bioinformatics: data's future shock. *Nature 428*, 6984 (Apr 2004), 774–777.

[39] Bulyk, M., Johnson, P., and Church, G. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res 30*, 5 (2002), 1255–1261.

[40] Bulyk, M. L., Huang, X., Choo, Y., and Church, G. M. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci U S A 98*, 13 (2001), 7158–7163.

[41] Bussemaker, H., Li, H., and Siggia, E. Regulatory element detection using correlation with expression. *Nat Genet 27*, 2 (2001), 167–171.

[42] Butcher, E. C. Can cell systems biology rescue drug discovery? *Nat Rev Drug Discov 4*, 6 (Jun 2005), 461–467.

[43] Butler, J. E. F., and Kadonaga, J. T. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev 16*, 20 (Oct 2002), 2583–2592.

[44] Case, D., Pearlman, D., Caldwell, J., Cheatham, T. I., Wang, J., Ross, W., Simmerling, C., Darden, T., Merz, K., Stanton, R., Cheng, A., Vincent, J., Crowley, M., Tsui, V., Gohlke, H., Radmer, R., Duan, Y., Pitera, J., Massova, I., Seibel, G., Singh, U., Weiner, P., and Kollman, P. *AMBER 7*. University of California, San Francisco, 2002.

[45] Cheng, L., Sturgis, E. M., Eicher, S. A., Char, D., Spitz, M. R., and Wei, Q. Glutathione-S-transferase polymorphisms and risk of squamous-cell carcinoma of the head and neck. *Int J Cancer 84*, 3 (Jun 1999), 220–224.

[46] Cheung, K. H., Nadkarni, P. M., and Shin, D. G. A metadata approach to query interoperation between molecular biology databases. *Bioinformatics 14*, 6 (1998), 486–497.

[47] Choo, Y. End effects in DNA recognition by zinc finger arrays. *Nucleic Acids Res 26*, 2 (1998), 554–557.

[48] Choo, Y., and Klug, A. Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc Natl Acad Sci U S A 91*, 23 (1994), 11168–11172.

[49] Choo, Y., and Klug, A. Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc Natl Acad Sci U S A 91*, 23 (1994), 11163–11167.

[50] Choo, Y., Sanchez-Garcia, I., and Klug, A. In vivo repression by a site-specific DNA-binding protein designed against an oncogene sequence. *Nature 372* (1994), 642–645.

[51] Clarke, N. D., and Berg, J. M. Zinc fingers in *Caenorhabditis elegans*: finding families and probing pathways. *Science 282*, 5396 (Dec 1998), 2018–2022.

[52] Comtesse, N., Heckel, D., Racz, A., Brass, N., Glass, B., and Meese, E. Five novel immunogenic antigens in meningioma: cloning, expression analysis, and chromosomal mapping. *Clin Cancer Res 5*, 11 (Nov 1999), 3560–3568.

[53] Cook, P. R. The organization of replication and transcription. *Science 284*, 5421 (Jun 1999), 1790–1795.

[54] Cora, D., Di Cunto, F., Provero, P., Silengo, L., and Caselle, M. Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs. *BMC Bioinformatics 5*, 1 (May 2004), 57.

[55] Cora, D., Herrmann, C., Dieterich, C., Di Cunto, F., Provero, P., and Caselle, M. Ab initio identification of putative human transcription factor binding sites by comparative genomics. *BMC Bioinformatics 6*, 1 (May 2005), 110.

[56] Cosma, M. P., Tanaka, T., and Nasmyth, K. Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. *Cell 97*, 3 (Apr 1999), 299–311.

[57] Darnell, J. E. J. Transcription factors as targets for cancer therapy. *Nat Rev Cancer 2*, 10 (Oct 2002), 740–749.

[58] Deng, Y., Glimm, J., Wang, Y., Korobka, A., Eisenberg, M., and Grollman, A. Prediction of protein binding to DNA in the presence of water-mediated hydrogen bonds. *J Mol Model 5*, 7-8 (1999), 125–133.

[59] Desjarlais, J. R., and Berg, J. M. Toward rules relating zinc finger protein sequences and DNA-binding site preferences. *Proc Natl Acad Sci U S A 89*, 16 (1992), 7345–7349.

[60] Desjarlais, J. R., and Berg, J. M. Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc Natl Acad Sci U S A 90*, 6 (Mar 1993), 2256–2260.

[61] Devaux, F., Carvajal, E., Moye-Rowley, S., and Jacq, C. Genome-wide studies on the nuclear PDR3-controlled response to mitochondrial dysfunction in yeast. *FEBS Lett 515*, 1-3 (Mar 2002), 25–28.

[62] Dickerson, R. E., and Chiu, T. K. Helix bending as a factor in protein-DNA recognition. *Biopolymers 44*, 4 (1997), 361–403.

[63] Dixit, S. B., Andrews, D. Q., and Beveridge, D. L. Induced fit and the entropy of structural adaptation in the complexation of CAP and lambda-repressor with cognate DNA sequences. *Biophys J 88*, 5 (May 2005), 3147–3157.

[64] Dong, C., Wilhelm, D., and Koopman, P. Sox genes and cancer. *Cytogenet Genome Res 105*, 2-4 (2004), 442–447.

[65] Dönnes, P., and Elofsson, A. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics 3*, 1 (Sep 2002), 25.

[66] Dönnes, P., and Höglund, A. Predicting protein subcellular localization: past, present, and future. *Genomics Proteomics Bioinformatics 2*, 4 (Nov 2004), 209–215.

[67] Dönnes, P., Höglund, A., Sturm, M., Comtesse, N., Backes, C., Meese, E., Kohlbacher, O., and Lenhof, H.-P. Integrative analysis of cancer-related data using CAP. *FASEB J 18*, 12 (Sep 2004), 1465–1467.

[68] Duret, L., and Bucher, P. Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol 7*, 3 (1997), 399–406.

[69] Elledge, S. J. Cell cycle checkpoints: preventing an identity crisis. *Science 274*, 5293 (Dec 1996), 1664–1672.

[70] Ellenberger, T. E., Brandl, C. J., Struhl, K., and Harrison, S. C. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. *Cell 71*, 7 (Dec 1992), 1223–1237.

[71] Elrod-Erickson, M., Rould, M., Nekludova, L., and Pabo, C. Zif268 protein-DNA complex refined at 1.6 å: a model system for understanding zinc finger-DNA interactions. *Structure 4*, 10 (1996), 1171–1180.

[72] Endres, R., Schulthess, T., and Wingreen, N. Toward an atomistic model for predicting transcription-factor binding sites. *Proteins 57*, 2 (2004), 262–268.

[73] Evans, J. N. S. *Biomolecular NMR Spectroscopy.* Oxford University Press, Oxford, US, 1995.

[74] Ferbus, D., Bovin, C., Validire, P., and Goubin, G. The zinc finger protein OZF (ZNF146) is overexpressed in colorectal cancer. *J Pathol 200*, 2 (Jun 2003), 177–182.

[75] Fessele, S., Maier, H., Zischek, C., Nelson, P. J., and Werner, T. Regulatory context is a crucial part of gene function. *Trends Genet 18*, 2 (Feb 2002), 60–63.

[76] Fickett, J., and Hatzigeorgiou, A. Eukaryotic promoter recognition. *Genome Res 7*, 9 (1997), 861–878.

[77] Fletcher, J. C., Brand, U., Running, M. P., Simon, R., and Meyerowitz, E. M. Signaling of cell fate decisions by CLAVATA3 in *Arabidopsis* shoot meristems. *Science 283*, 5409 (Mar 1999), 1911–1914.

[78] Fogolari, F., Brigo, A., and Molinari, H. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J Mol Recognit 15*, 6 (Nov 2002), 377–392.

[79] Franklin, R. E., and Gosling, R. G. Molecular configuration in sodium thymonucleate. 1953. *Nature 421*, 6921 (Jan 2003), 400–401.

[80] Frech, K., Herrmann, G., and Werner, T. Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res 21*, 7 (1993), 1655–1664.

[81] Frech, K., and Werner, T. Specific modelling of regulatory units in DNA sequences. *Pac Symp Biocomput* (1997), 151–162.

[82] Frith, M., Hansen, U., and Weng, Z. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics 17*, 10 (2001), 878–889.

[83] Fry, C. J., and Farnham, P. J. Context-dependent transcriptional regulation. *J Biol Chem 274*, 42 (Oct 1999), 29583–29586.

[84] Galas, D., and Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res 5*, 9 (1978), 3157–3170.

[85] Galibert, M. D., Carreira, S., and Goding, C. R. The Usf-1 transcription factor is a novel target for the stress-responsive p38 kinase and mediates UV-induced Tyrosinase expression. *EMBO J 20*, 17 (Sep 2001), 5022–5031.

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

[87] Gasch, A. P., Huang, M., Metzner, S., Botstein, D., Elledge, S. J., and Brown, P. O. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol Biol Cell 12*, 10 (Oct 2001), 2987–3003.

[88] Getz, G., Levine, E., Domany, E., and Zhang, M. Q. Super-paramagnetic clustering of yeast gene expression profiles. *Physica A 279* (2000), 457–464.

[89] Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A. P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.-D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kotter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C.-y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W., and Johnston, M. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature 418*, 6896 (Jul 2002), 387–391.

[90] Gibson, G. Microarray analysis: genome-scale hypothesis scanning. *PLoS Biol 1*, 1 (Oct 2003), E15.

[91] Gohlke, H., and Case, D. A. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J Comput Chem 25*, 2 (Jan 2004), 238–250.

[92] Gohlke, H., and Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew Chem Int Ed Engl 41*, 15 (Aug 2002), 2644–2676.

[93] Gorfe, A., Caflisch, A., and Jelesarov, I. The role of flexibility and hydration on the sequence-specific DNA recognition by the Tn916 integrase protein: a molecular dynamics analysis. *J Mol Recognit 17*, 2 (2004), 120–131.

[94] Greisman, H. A., and Pabo, C. O. A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science 275*, 5300 (Jan 1997), 657–661.

[95] Gromiha, M., Siebers, J., Selvaraj, S., Kono, H., and Sarai, A. Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J Mol Biol 337*, 2 (2004), 285–294.

[96] Grundy, W. N., Bailey, T. L., and Elkan, C. P. ParaMEME: a parallel implementation and a web interface for a DNA and protein motif discovery tool. *Comput Appl Biosci 12*, 4 (Aug 1996), 303–310.

[97] GuhaThakurta, D., and Stormo, G. D. Identifying target sites for cooperatively binding factors. *Bioinformatics 17*, 7 (2001), 608–621.

[98] Hannenhalli, S., and Levy, S. Promoter prediction in the human genome. *Bioinformatics 17 Suppl 1* (2001), 90–96.

[99] Hannenhalli, S., and Levy, S. Predicting transcription factor synergism. *Nucleic Acids Res 30*, 19 (2002).

[100] Haverty, P. M., Hansen, U., and Weng, Z. Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Res 32*, 1 (2004), 179–188.

[101] Hazan, C., Melzer, K., Panageas, K. S., Li, E., Kamino, H., Kopf, A., Cordon-Cardo, C., Osman, I., and Polsky, D. Evaluation of the proliferation marker MIB-1 in the prognosis of cutaneous malignant melanoma. *Cancer 95*, 3 (Aug 2002), 634–640.

[102] Helms, V., and Wade, R. Computational alchemy to calculate absolute protein-ligand binding free energy. *J. Am. Chem. Soc. 120* (1998), 2710–2713.

[103] Henrique, R., Azevedo, R., Bento, M. J., Domingues, J. C., Silva, C., and Jeronimo, C. Prognostic value of Ki-67 expression in localized cutaneous malignant melanoma. *J Am Acad Dermatol 43*, 6 (Dec 2000), 991–1000.

[104] Höglund, A., Dönnes, P., Blum, T., Adolph, H.-W., and Kohlbacher, O. Using N-terminal targeting sequences, amino acid composition, and sequence motifs for predicting protein subcellular localization. *Proceedings of the German Conference on Bioinformatics (GCB 2005), edited by Andrew Torda, Stefan Kurtz, Matthias Rarey* (2005), 45–59.

[105] Höglund, A., and Kohlbacher, O. From sequence to structure and back again: approaches for predicting protein-DNA binding. *Proteome Sci 2*, 1 (Jun 2004), 3.

[106] Hong, R. L., Hamaguchi, L., Busch, M. A., and Weigel, D. Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. *Plant Cell 15*, 6 (Jun 2003), 1296–1309.

[107] Hsu, S.-H., Hsieh-Li, H.-M., Huang, H.-Y., Huang, P.-H., and Li, H. bHLH-zip transcription factor Spz1 mediates mitogen-activated protein kinase cell proliferation, transformation, and tumorigenesis. *Cancer Res 65*, 10 (May 2005), 4041–4050.

[108] Hua, S., and Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics 17*, 8 (Aug 2001), 721–728.

[109] Huang, M., Zhou, Z., and Elledge, S. J. The DNA replication and damage checkpoint pathways induce transcription by inhibition of the Crt1 repressor. *Cell 94*, 5 (Sep 1998), 595–605.

[110] Hughes, J., Estep, P., Tavazoie, S., and Church, G. Computational identification of cis-regulatory elements associated with functionally coherent groups of genes in *Saccharomyces cerevisiae*. *J Mol Biol 296*, 5 (2000), 1205–1214.

[111] Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics 18 Suppl 1* (2002), 233–240.

[112] Ideker, T., Thorsson, V., Ranish, J., Christmas, R., Buhler, J., Eng, J., Bumgarner, R., Goodlett, D., Aebersold, R., and Hood, L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science 292*, 5518 (2001), 929–934.

[113] Isalan, M., Choo, Y., and Klug, A. Synergy between adjacent zinc fingers in sequence-specific DNA recognition. *Proc Natl Acad Sci U S A 94*, 11 (1997), 5617–5621.

[114] Isalan, M., Klug, A., and Choo, Y. Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers. *Biochemistry 37*, 35 (1998), 12026–12033.

[115] Isalan, M., Klug, A., and Choo, Y. A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nat Biotechnol 19*, 7 (Jul 2001), 656–660.

[116] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A 98*, 8 (Apr 2001), 4569–4574.

[117] Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature 409*, 6819 (Jan 2001), 533–538.

[118] Jamieson, A. C., Miller, J. C., and Pabo, C. O. Drug discovery with engineered zinc-finger proteins. *Nat Rev Drug Discov 2*, 5 (May 2003), 361–368.

[119] Jayaram, B., and Jain, T. The role of water in protein-DNA recognition. *Annual Review of Biophysics and Biomolecular Structure 33* (2004), 343–361.

[120] Jayaram, B., McConnell, K., Dixit, S. B., Das, A., and Beveridge, D. L. Free-energy component cnalysis of 40 protein-DNA complexes: a consensus view on the thermodynamics of binding at the molecular level. *J Comput Chem 23*, 1 (2002), 1–14.

[121] Jensen, L. J., Gupta, R., Staerfeldt, H.-H., and Brunak, S. Prediction of human protein function according to Gene Ontology categories. *Bioinformatics 19*, 5 (Mar 2003), 635–642.

[122] Johansson, O., Alkema, W., Wasserman, W., and Lagergren, J. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics 19*, Suppl 1 (2003), 169–176.

[123] Jones, S., van Heyningen, P., Berman, H., and Thornton, J. Protein-DNA interactions: A structural analysis. *J Mol Biol 287*, 5 (1999), 877–896.

[124] Jonsson, H., Heisler, M., Reddy, G. V., Agrawal, V., Gor, V., Shapiro, B. E., Mjolsness, E., and Meyerowitz, E. M. Modeling the organization of the WUSCHEL expression domain in the shoot apical meristem. *Bioinformatics 21 Suppl 1* (Jun 2005), i232–i240.

[125] Jorgensen, W. L., Chandresekhar, J., Madura, J., Impey, R., and Klein, M. Comparison of simple potential functions for simulating liquid water. *J Chem Phys 79* (1983), 926–935.

[126] Karas, H., Knuppel, R., Schulz, W., Sklenar, H., and Wingender, E. Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. *Comput Appl Biosci 12*, 5 (1996), 441–446.

[127] Karplus, M., and Petsko, G. A. Molecular dynamics simulations in biology. *Nature 347*, 6294 (Oct 1990), 631–639.

[128] Kielbasa, S., Korbel, J., Beule, D., Schuchhardt, J., and Herzel, H. Combining frequency and positional information to predict transcription factor binding sites. *Bioinformatics 17*, 11 (2001), 1019–1026.

[129] Kielkopf, C. L., White, S., Szewczyk, J. W., Turner, J. M., Baird, E. E., Dervan, P. B., and Rees, D. C. A structural basis for recognition of A.T and T.A base pairs in the minor groove of B-DNA. *Science 282*, 5386 (Oct 1998), 111–115.

[130] Kim, J., and Burley, S. 1.9 A resolution refined structure of TBP recognizing the minor groove of TATAAAAG. *Nat Struct Biol 1*, 9 (1994), 638–653.

[131] Kim, J., Tzamarias, D., Ellenberger, T., Harrison, S. C., and Struhl, K. Adaptability at the protein-DNA interface is an important aspect of sequence recognition by bZIP proteins. *Proc Natl Acad Sci U S A 90*, 10 (May 1993), 4513–4517.

[132] Kim, J. S., and Pabo, C. O. Getting a handhold on DNA: design of poly-zinc finger proteins with femtomolar dissociation constants. *Proc Natl Acad Sci U S A 95*, 6 (Mar 1998), 2812–2817.

[133] King, O. D., and Roth, F. P. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res 31*, 19 (Oct 2003), e116.

[134] Kise, K. J. J., and Shin, J. A. The contribution of the methyl groups on thymine bases to binding specificity and affinity by alanine-rich mutants of the bZIP motif. *Bioorg Med Chem 9*, 9 (Sep 2001), 2485–2491.

[135] Kitano, H. Computational systems biology. *Nature 420*, 6912 (Nov 2002), 206–210.

[136] Kitano, H. Systems biology: a brief overview. *Science 295*, 5560 (Mar 2002), 1662–1664.

[137] Klug, A., and Rhodes, D. Zinc fingers: a novel protein fold for nucleic acid recognition. *Cold Spring Harb Symp Quant Biol 52* (1987), 473–482.

[138] Kohler, J. J., and Schepartz, A. Kinetic studies of Fos.Jun.DNA complex formation: DNA binding prior to dimerization. *Biochemistry 40*, 1 (Jan 2001), 130–142.

[139] Kollman, P., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D., and Cheatham, T. r. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res 33*, 12 (2000), 889–897.

[140] Kollman, P. A., Weiner, S., Seibel, G., Lybrand, T., Singh, U. C., Caldwell, J., and Rao, S. N. Modeling complex molecular interactions involving proteins and DNA. *Ann N Y Acad Sci 482* (1986), 234–244.

[141] Kono, H., and Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins 35*, 1 (1999), 114–131.

[142] Krivan, W., and Wasserman, W. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res 11*, 9 (2001), 1559–1566.

[143] Krizek, B., Amann, B. T., Kilfoil, V. J., Merkle, D. L., and Berg, J. A Consensus Zinc Finger Peptide: Design, High Affinity Metal Binding, pH Dependent Structure, and a His to Cys Sequence Variant. *J. Am. Chem. Soc. 113* (1991), 4518–4523.

[144] Ladomery, M., and Dellaire, G. Multifunctional zinc finger proteins in development and disease. *Ann Hum Genet 66*, Pt 5-6 (Nov 2002), 331–342.

[145] Laux, T., Mayer, K. F., Berger, J., and Jurgens, G. The WUSCHEL gene is required for shoot and floral meristem integrity in *Arabidopsis*. *Development 122*, 1 (Jan 1996), 87–96.

[146] Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A., and Wootton, J. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science 262*, 5131 (1993), 208–214.

[147] Lazaridis, T. Binding affinity and specificity from computational studies. *Current Organic Chemistry 6* (2002), 1319–1332.

[148] Leach, A. *Molecular Modelling. Principles and Applications, 2nd Edition.* Pearson Education Limited, Esses, UK, 2001.

[149] Lee, M., Bulyk, M., Whitmore, G., and Church, G. A statistical model for investigating binding probabilities of DNA nucleotide sequences using microarrays. *Biometrics 58*, 4 (2002), 981–988.

[150] Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science 298*, 5594 (Oct 2002), 799–804.

[151] Lee, Dong-ki Seol, W., and Kim, J.-S. Custom dna-binding proteins and artificial transcription factors. *Current Topics in Medicinal Chemistry 3* (2003), 339–353.

[152] Lemon, W., Liyanarachchi, S., and You, M. A high performance test of differential gene expression for oligonucleotide arrays. *Genome Biol. 4*, 10 (Epub 2003).

[153] Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N., and Wasserman, W. Identification of conserved regulatory elements by comparative genome analysis. *J Biol 2*, 2 (2003), Epub.

[154] Lenhard, B., and Wasserman, W. W. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics 18*, 8 (Aug 2002), 1135–1136.

[155] Leslie, M. C., and Bar-Eli, M. Regulation of gene expression in melanoma: new approaches for treatment. *J Cell Biochem 94*, 1 (Jan 2005), 25–38.

[156] Levine, M., and Davidson, E. H. Gene regulatory networks for development. *Proc Natl Acad Sci U S A 102*, 14 (Apr 2005), 4936–4942.

[157] Levy, S., and Hannenhalli, S. Identification of transcription factor binding sites in the human genome sequence. *Mamm Genome 13*, 9 (2002), 510–514.

[158] Levy, S., Hannenhalli, S., and Workman, C. Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics 17*, 10 (2001), 871–877.

[159] Liu, Q., Segal, D. J., Ghiara, J. B., and Barbas, C. F. r. Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. *Proc Natl Acad Sci U S A 94*, 11 (May 1997), 5525–5530.

[160] Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol 14*, 13 (Dec 1996), 1675–1680.

[161] Lohmann, J. U., Hong, R. L., Hobe, M., Busch, M. A., Parcy, F., Simon, R., and Weigel, D. A molecular link between stem cell regulation and floral patterning in Arabidopsis. *Cell 105*, 6 (Jun 2001), 793–803.

[162] Long, J. A., and Barton, M. K. The development of apical embryonic pattern in *Arabidopsis*. *Development 125*, 16 (Aug 1998), 3027–3035.

[163] Longabaugh, W. J. R., Davidson, E. H., and Bolouri, H. Computational representation of developmental genetic regulatory networks. *Dev Biol 283*, 1 (Jul 2005), 1–16.

[164] Lowney, J. K., Boucher, L. D., Swanson, P. E., and Doherty, G. M. Interferon regulatory factor-1 and -2 expression in human melanoma specimens. *Ann Surg Oncol 6*, 6 (Sep 1999), 604–608.

[165] Luscombe, N., Austin, S., Berman, H., and Thornton, J. An overview of the structures of protein-DNA complexes. *Genome Biol 1*, 1 (2000), Epub, Review.

[166] Luscombe, N. M., Laskowski, R. A., and Thornton, J. M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res 29*, 13 (2001), 2860–2874.

[167] Lynch, M., and Conery, J. S. The origins of genome complexity. *Science 302*, 5649 (Nov 2003), 1401–1404.

[168] Mandel-Gutfreund, Y., Baron, A., and Margalit, H. A structure-based approach for prediction of protein binding sites in gene upstream regions. *Pac Symp Biocomput* (2001), 139–150.

[169] Mandel-Gutfreund, Y., Schueler, O., and Margalit, H. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J Mol Biol 253*, 2 (1995), 370–382.

[170] Manning, G. The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Q. Rev. Biophys. 11* (1978), 179–246.

[171] Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct 29* (2000), 291–325.

[172] Martinez, J. A., Prevot, S., Nordlinger, B., Nguyen, T. M., Lacarriere, Y., Munier, A., Lascu, I., Vaillant, J. C., Capeau, J., and Lacombe, M. L. Overexpression of nm23-H1 and nm23-H2 genes in colorectal carcinomas and loss of nm23-H1 expression in advanced tumour stages. *Gut 37*, 5 (Nov 1995), 712–720.

[173] Matthews, B. Protein-DNA interaction. No code for recognition. *Nature 335*, 6188 (1988), 294–295.

[174] Mayer, K. F., Schoof, H., Haecker, A., Lenhard, M., Jurgens, G., and Laux, T. Role of WUSCHEL in regulating stem cell fate in the *Arabidopsis* shoot meristem. *Cell 95*, 6 (Dec 1998), 805–815.

[175] McCammon, J. A. Theory of biomolecular recognition. *Curr Opin Struct Biol 8* (1998), 245–249.

[176] Meese, E., and Comtesse, N. Cancer genetics and tumor antigens: time for a combined view? *Genes Chromosomes Cancer 33*, 2 (Feb 2002), 107–113.

[177] Miller, J., and Pabo, C. Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. *J Mol Biol 313*, 2 (2001), 309–315.

[178] Misra, V. K., Hecht, J. L., Sharp, K. A., Friedman, R. A., and Honig, B. Salt effects on protein-DNA interactions. The lambda cI repressor and EcoRI endonuclease. *J Mol Biol 238*, 2 (Apr 1994), 264–280.

[179] Mitrunen, K., and Hirvonen, A. Molecular epidemiology of sporadic breast cancer. The role of polymorphic genes involved in oestrogen biosynthesis and metabolism. *Mutat Res 544*, 1 (Sep 2003), 9–41.

[180] Moore, M., Klug, A., and Choo, Y. Improved DNA binding specificity from polyzinc finger peptides by using strings of two-finger units. *Proc Natl Acad Sci U S A 98*, 4 (2001), 1437–1436.

[181] Mossner, R., Schulz, U., Kruger, U., Middel, P., Schinner, S., Fuzesi, L., Neumann, C., and Reich, K. Agonists of peroxisome proliferator-activated receptor gamma inhibit cell growth in malignant melanoma. *J Invest Dermatol 119*, 3 (Sep 2002), 576–582.

[182] Nakai, K., and Kanehisa, M. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics 14*, 4 (Dec 1992), 897–911.

[183] Nardelli, J., Gibson, T., and Charnay, P. Zinc finger-DNA recognition: analysis of base specificity by site-directed mutagenesis. *Nucleic Acids Res 20*, 16 (Aug 1992), 4137–4144.

[184] Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol 48*, 3 (Mar 1970), 443–453.

[185] Nishie, A., Masuda, K., Otsubo, M., Migita, T., Tsuneyoshi, M., Kohno, K., Shuin, T., Naito, S., Ono, M., and Kuwano, M. High expression of the Cap43 gene in infiltrating macrophages of human renal cell carcinomas. *Clin Cancer Res 7*, 7 (Jul 2001), 2145–2151.

[186] Norman, C., Runswick, M., Pollock, R., and Treisman, R. Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the c-fos serum response element. *Cell 55*, 6 (Dec 1988), 989–1003.

[187] Nowak, R. Mining treasures from 'junk DNA'. *Science 263*, 5147 (Feb 1994), 608–610. News.

[188] Pabo, C., and Sauer, R. Protein-DNA recognition. *Annu Rev Biochem 53* (1984), 293–321.

[189] Pabo, C. O., and Nekludova, L. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J Mol Biol 301*, 3 (2000), 597–624.

[190] Pabo, C. O., and Sauer, R. T. Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem 61* (1992), 1053–1095.

[191] Paillard, G., and Lavery, R. Analyzing protein-DNA recognition mechanisms. *Structure (Camb) 12*, 1 (Jan 2004), 113–122.

[192] Pavletich, N. P., and Pabo, C. O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 å. *Science 252*, 5007 (1991), 809–817.

[193] Pe'er, D., Regev, A., Elidan, G., and Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics 17 Suppl 1* (2001), 215–224.

[194] Pennisi, E. Bioinformatics. Gene counters struggle to get the right answer. *Science 301*, 5636 (Aug 2003), 1040–1041.

[195] Pichierri, F., Aida, M., Gromiha, M., and Sarai, A. Free-energy maps of base-amino acid interactions for protein-DNA recognition. *J Am Chem Soc 121* (1999), 6152–6257.

[196] Pietrokovski, S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res 24*, 19 (Oct 1996), 3836–3845.

[197] Pilpel, Y., Sudarsanam, P., and Church, G. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet 29*, 2 (2001), 153–159.

[198] Pollard, T. D., and Earnshaw, W. C. *Cell Biology.* WB Saunders/Elsevier, Philadelphia, US, 2004.

[199] Pomerantz, J. L., Sharp, P. A., and Pabo, C. O. Structure-based design of transcription factors. *Science 267*, 5194 (Jan 1995), 93–96.

[200] Pruitt, K. D., and Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res 29*, 1 (Jan 2001), 137–140.

[201] Qiu, P. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochem Biophys Res Commun 309*, 3 (Sep 2003), 495–501.

[202] Ravishanker, G., Auffinger, P., Langley, D. R., Jayaram, B., Young, M. A., and Beveridge, D. L. Treatment of counterions in computer simulations of DNA. In *Reviews in Comp. Chem.*, K. B. Lipkovitz and D. B. Boyd, Eds., vol. 11. Wiley-VCH, New York, 1997, pp. 317–372.

[203] Rebar, E. J., Huang, Y., Hickey, R., Nath, A. K., Meoli, D., Nath, S., Chen, B., Xu, L., Liang, Y., Jamieson, A. C., Zhang, L., Spratt, S. K., Case, C. C., Wolffe, A., and Giordano, F. J. Induction of angiogenesis in a mouse model using engineered transcription factors. *Nat Med 8*, 12 (Dec 2002), 1427–1432.

[204] Reddy, C. K., Das, A., and Jayaram, B. Do water molecules mediate protein-DNA recognition? *J Mol Biol 314*, 3 (Nov 2001), 619–632.

[205] Reddy, M. R., and Erion, M. *Free Energy Calculations in Rational Drug Design.* Kluwer Academic/Plenum Publishers, New York, US, 2001.

[206] Reddy, M. R., Erion, M., and Agarwal, A. Free Energy Calculations: Use and Limitations in Predicting Ligand Binding Affinity. In *Reviews in computational chemistry*, K. B. Lipkowitz and D. B. Boyd, Eds., vol. 16. Wiley-VCH Inc., New York, US, 2000.

[207] Reinhardt, D., Frenz, M., Mandel, T., and Kuhlemeier, C. Microsurgical and laser ablation analysis of interactions between the zones and layers of the tomato shoot apical meristem. *Development 130*, 17 (Sep 2003), 4073–4083.

[208] Remacle, J. E., Kraft, H., Lerchner, W., Wuytens, G., Collart, C., Verschueren, K., Smith, J. C., and Huylebroeck, D. New mode of DNA binding of multi-zinc finger transcription factors: deltaEF1 family members bind with two hands to two target sites. *EMBO J 18*, 18 (Sep 1999), 5073–5084.

[209] Ren, B., Robert, F., Wyrick, J., Aparicio, O., Jennings, E., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T., Wilson, C., Bell, S., and Young, R. Genome-wide location and function of DNA binding proteins. *Science 290*, 5500 (2000), 2306–2309.

[210] Rhodes, G. *Crystallography Made Crystal Clear, 2nd Edition.* Academic Press, London, 1999.

[211] Ribeiro, R. C., Kushner, P. J., and Baxter, J. D. The nuclear hormone receptor gene superfamily. *Annu Rev Med 46* (1995), 443–453.

[212] Roth, F. P., Hughes, J. D., Estep, P. W., and Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol 16*, 10 (Oct 1998), 939–945.

[213] Roulet, E., Busso, S., Camargo, A., Simpson, A., Mermod, N., and Bucher, P. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol 20*, 8 (2002), 831–835.

[214] Rumbauugh, J., Jacobson, I., and Booch, G. *The Unified Modelling Launguage Reference Manual.* Addison-Wesley, New York, USA, 1999.

[215] Saito, M., and Sarai, K. Free energy calculations for the relative binding affinity between DNA and lambda-repressor. *Proteins 52*, 2 (2003), 129–136.

[216] Sanchez, J.-P., Ullman, C., Moore, M., Choo, Y., and Chua, N.-H. Regulation of gene expression in *Arabidopsis thaliana* by artificial zinc finger chimeras. *Plant Cell Physiol 43*, 12 (Dec 2002), 1465–1472.

[217] Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., and Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res 32*, Database issue (Jan 2004), 91–94.

[218] Sandelin, A., Höglund, A., Lenhard, B., and Wasserman, W. Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes. *Funct Integr Genomics 3*, 3 (2003), 125–134.

[219] Sarai A, T. Y. Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. *Proc Natl Acad Sci U S A 86*, 17 (1989), 6513–6517.

[220] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science 270*, 5235 (Oct 1995), 467–470.

[221] Schlick, T. *Molecular Modeling and Simulation.* Springer, New York, US, 2002.

[222] Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J Mol Biol 188*, 3 (Apr 1986), 415–431.

[223] Schoof, H., Lenhard, M., Haecker, A., Mayer, K. F., Jurgens, G., and Laux, T. The stem cell population of *Arabidopsis* shoot meristems in maintained by a regulatory loop between the CLAVATA and WUSCHEL genes. *Cell 100*, 6 (Mar 2000), 635–644.

[224] Schutz, C. N., and Warshel, A. What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins 44*, 4 (Sep 2001), 400–417.

[225] Schwabe, J. The role of water in protein-DNA interactions. *Curr Opin Struct Biol 7*, 1 (1997), 126–134.

[226] Seeman NC, Rosenberg JM, R. A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U S A 73*, 3 (1976), 804–808.

[227] Selvaraj, S., Kono, H., and Sarai, A. Specificity of protein-DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding. *J Mol Biol 322*, 5 (2002), 907–915.

[228] Sen, S., and Nilsson, L. Free energy calculations and molecular dynamics simulations of wild-type and variants of the DNA-EcoRI complex. *Biophys J 77*, 4 (1999), 1801–1810.

[229] Shannon, M. F., and Rao, S. Transcription. Of chips and ChIPs. *Science 296*, 5568 (Apr 2002), 666–669.

[230] Sharma, V. K., Carles, C., and Fletcher, J. C. Maintenance of stem cell populations in plants. *Proc Natl Acad Sci U S A 100 Suppl 1* (Sep 2003), 11823–11829.

[231] Shimizu, T., Abe, R., Nakamura, H., Ohkawara, A., Suzuki, M., and Nishihira, J. High expression of macrophage migration inhibitory factor in human melanoma cells and its role in tumor cell growth and angiogenesis. *Biochem Biophys Res Commun 264*, 3 (Nov 1999), 751–758.

[232] Shiraishi, H., Okada, K., and Shimura, Y. Nucleotide sequences recognized by the AGAMOUS MADS domain of *Arabidopsis thaliana* in vitro. *Plant J 4*, 2 (Aug 1993), 385–398.

[233] Siggers, T. W., Silkov, A., and Honig, B. Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. *J Mol Biol 345*, 5 (Feb 2005), 1027–1045.

[234] Soussi, T. p53 Antibodies in the sera of patients with various types of cancer: a review. *Cancer Res 60*, 7 (Apr 2000), 1777–1788.

[235] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell 9*, 12 (Dec 1998), 3273–3297.

[236] St Johnston, D., and Nüsslein-Volhard, C. The origin of pattern and polarity in the Drosophila embryo. *Cell 68*, 2 (Jan 1992), 201–219.

[237] Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res 12*, 10 (Oct 2002), 1611–1618.

[238] Stapley, B. J., Kelley, L. A., and Sternberg, M. J. E. Predicting the sub-cellular location of proteins from text using support vector machines. *Pac Symp Biocomput* (2002), 374–385.

[239] Steeves, T. A., and Sussex, I. M. *Patterns in Plant Development.* Cambridge University Press, Cambridge, 1989.

[240] Steffen, N., Murphy, S., Tolleri, L., Hatfield, G., and Lathrop, R. DNA sequence and structure: direct and indirect recognition in protein-DNA binding. *Bioinformatics 18*, 1 (2002), 522–530.

[241] Steffen, N. R., Murphy, S. D., Lathrop, R. H., Opel, M. L., Tolleri, L., and Hatfield, G. W. The role of DNA deformation energy at individual base steps for the identification of DNA-protein binding sites. *Genome Inform Ser Workshop Genome Inform 13* (2002), 153–162.

[242] Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J. r., and Doyle, J. Robustness of cellular functions. *Cell 118*, 6 (Sep 2004), 675–685.

[243] Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc. 112* (1990), 6127–6129.

[244] Stormo, G. DNA binding sites: representation and discovery. *Bioinformatics 16*, 1 (2000), 16–23.

[245] Stormo, G. D. Consensus patterns in DNA. *Methods Enzymol 183* (1990), 211–221.

[246] Struhl, K. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell 98*, 1 (Jul 1999), 1–4.

[247] Stryer, L. *Biochemistry, 4th Edition.* W.H. Freeman and Company, New York, US, 1995.

[248] Suzuki, M., Brenner, S., Gerstein, M., and Yagi, N. DNA recognition code of transcription factors. *Protein Eng 8*, 4 (1995), 319–328.

[249] Swirnoff, A. H., and Milbrandt, J. DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol Cell Biol 15*, 4 (Apr 1995), 2275–2287.

[250] Tellez, C., McCarty, M., Ruiz, M., and Bar-Eli, M. Loss of activator protein-2alpha results in overexpression of protease-activated receptor-1 and correlates with the malignant phenotype of human melanoma. *J Biol Chem 278*, 47 (Nov 2003), 46632–46642.

[251] Thieffry, D., Huerta, A. M., Perez-Rueda, E., and Collado-Vides, J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays 20*, 5 (May 1998), 433–440.

[252] Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol 9*, 2 (2002), 447–464.

[253] Thompson, J. D., Higgins, D. G., and Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res. 22*, 22 (1994), 4673–4680.

[254] Tong, W. Analyzing the biology on the system level. *Genomics Proteomics Bioinformatics 2*, 1 (Feb 2004), 6–14.

[255] Travers, A. A. DNA bending by sequence and proteins. In *DNA-Protein: Structural Interactions*, D. Lilley, Ed. IRL Press, Oxford, 1995, pp. 49–75.

[256] Tsui, V., Radhakrishnan, I., Wright, P., and Case, D. NMR and molecular dynamics studies of the hydration of a zinc finger-DNA complex. *J Mol Biol 302*, 5 (2000), 1101–1117.

[257] Tureci, O., Sahin, U., and Pfreundschuh, M. Serological analysis of human tumor antigens: molecular definition and implications. *Mol Med Today 3*, 8 (Aug 1997), 342–349.

[258] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature 403*, 6770 (Feb 2000), 623–627.

[259] Urnov, F. D., and Wolffe, A. P. Chromatin remodeling and transcriptional activation: the cast (in order of appearance). *Oncogene 20*, 24 (May 2001), 2991–3006.

[260] van Helden, J., Andre, B., and Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol 281*, 5 (Sep 1998), 827–842.

[261] van Helden, J., Andre, B., and Collado-Vides, J. A web site for the computational analysis of yeast regulatory sequences. *Yeast 16*, 2 (Jan 2000), 177–187.

[262] Varnai, P., and Zakrzewska, K. DNA and its counterions: a molecular dynamics study. *Nucleic Acids Res 32*, 14 (2004), 4269–4280.

[263] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S.,

Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. The sequence of the human genome. *Science 291*, 5507 (Feb 2001), 1304–1351.

[264] von Hippel, P. H., and Berg, O. G. On the specificity of DNA-protein interactions. *Proc Natl Acad Sci U S A 83*, 6 (1986), 1608–1612.

[265] Waites, R., and Simon, R. Signaling cell fate in plant meristems. Three clubs on one tousle. *Cell 103*, 6 (Dec 2000), 835–838.

[266] Wang, J., Cieplak, P., and Kollman, P. How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem 21*, 12 (2000), 1049–1074.

[267] Wasserman, W., and Fickett, J. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol 278*, 1 (1998), 167–181.

[268] Wasserman, W., Palumbo, M., Thompson, W., Fickett, J., and Lawrence, C. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet 26*, 2 (2000), 225–228.

[269] Watanabe, K., and Okada, K. Two discrete cis elements control the Abaxial side-specific expression of the FILAMENTOUS FLOWER gene in *Arabidopsis*. *Plant Cell 15*, 11 (Nov 2003), 2592–2602.

[270] Watson, J., and Crick, F. The structure of DNA. *Cold Spring Harb Symp Quant Biol 18* (1953), 123–131.

[271] Weigel, D., and Jurgens, G. Stem cells that make stems. *Nature 415*, 6873 (Feb 2002), 751–754.

[272] Weng, G., Bhalla, U. S., and Iyengar, R. Complexity in biological signaling systems. *Science 284*, 5411 (Apr 1999), 92–96.

[273] Werner, E. In silico multicellular systems biology and minimal genomes. *Drug Discov Today 8*, 24 (Dec 2003), 1121–1127.

[274] Werner, T. Models for prediction and recognition of eukaryotic promoters. *Mamm Genome 10*, 2 (1999), 168–175.

[275] Werner, T. The state of the art of mammalian promoter recognition. *Brief Bioinform 4*, 1 (2003), 22–30.

[276] White, S., Khaliq, F., Sotiriou, S., and McInerny, C. J. The role of DSC1 components cdc10+, rep1+ and rep2+ in MCB gene transcription at the mitotic G1-S boundary in fission yeast. *Curr Genet 40*, 4 (Dec 2001), 251–259.

[277] WILKINS, M. H. F., STOKES, A. R., and WILSON, H. R. Molecular structure of deoxypentose nucleic acids. *Nature 171*, 4356 (Apr 1953), 738–740.

[278] Woda, J., Schneider, B., Patel, K., Mistry, K., and Berman, H. An analysis of the relationship between hydration and protein-DNA interactions. *Biophys J 75*, 5 (1998), 2170–2177.

[279] Wolfe, S. A., Grant, R. A., Elrod-Erickson, M., and Pabo, C. O. Beyond the "recognition code": structures of two Cys2His2 zinc finger/TATA box complexes. *Structure (Camb) 9*, 8 (Aug 2001), 717–723.

[280] Wolfe, S. A., Greisman, H. A., Ramm, E. I., and Pabo, C. O. Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J Mol Biol 285*, 5 (Feb 1999), 1917–1934.

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

[282] Workman, C. T., Yin, Y., Corcoran, D. L., Ideker, T., Stormo, G. D., and Benos, P. V. enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res 33*, Web Server issue (Jul 2005), 389–392.

[283] Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *Johns Hopkins University, Dept. of Biostatistics Working Papers.*, 1 (2004).

[284] Xing, E., Wu, W., Jordan, M., and Karp, R. LOGOS: a modular Bayesian model for de novo motif detection. *IEEE Computer Society Bioinformatics Conference, CSB2003* (2003).

[285] Yoshida, T., Nishimura, T., Aida, M., Pichierri, F., and Gromiha, M. M. n. S. A. Evaluation of free energy landscape for base-amino acid interactions using ab initio force field and extensive sampling. *Biopolymers 61*, 1 (2001-2002), 84–95.

[286] Yuan, D., Ma, X., and Ma, J. Recognition of multiple patterns of DNA sites by Drosophila homeodomain protein Bicoid. *J Biochem (Tokyo) 125*, 4 (Apr 1999), 809–817.

[287] Zhang, L. Y., Gallicchio, E., Friesner, R. A., and Levy, R. M. Solvent models for protein-ligand binding: Comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations. *J Comp Chem 22*, 6 (2001), 591–607.

[288] Zhang, Z., and Gerstein, M. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J Biol 2*, 2 (2003), Epub.

[289] Zhou, Y., Luoh, S.-M., Zhang, Y., Watanabe, C., Wu, T. D., Ostland, M., Wood, W. I., and Zhang, Z. Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis. *Cancer Res 63*, 18 (Sep 2003), 5781–5784.

[290] Zhu, J., and Zhang, M. Q. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics 15*, 7-8 (Jul 1999), 607–611.

The hardest part is knowing when you are finished

Unknown Artist