On the characterization of protein-DNA interactions using statistical potentials and protein-protein interactions

Oriol Fornés Crespo

DOCTORAL THESIS UPF / 2014

THESIS DIRECTOR:

Dr. Baldomero Oliva Miguel

Structural Bioinformatics Lab (SBI)
Research Program on Biomedical Informatics (GRIB)
Department of Experimental and Health Sciences (CEXS)

upf.
**Universitat
Pompeu Fabra**
Barcelona

Als meus pares,
per tot el seu suport i la seva paciencia.
Sense vosaltres això no hauria estat possible.

Acknowledgements

This thesis was supported by the FEDER BIO2011-22568 grant from the Spanish Ministry of Science and Innovation (MICINN).

I'd like to start by thanking my thesis supervisor, Prof. Baldo Oliva, who is at least as responsible of this work as I am myself. It has always been a pleasure to work under your supervision, which I deeply admire. Thank you Baldo for giving me the chance to make a doctorate in your group and for always seeing things positively and with an open mind. It is an inspiration to learn from you.

This thesis wouldn't be the same either without my SBI fellow lab mates. Your contribution in the form of science and/or friendship has been priceless. Thanks Jordi, Ramón, Dani, Joan, David, Jaume, Javi, Aggeliki, Emre, Olga, Elisenda, Manuel, Daniel, Alessandra, Elin, Attila, Billur, and Bernat for sharing much more than just the office with me. All your names well deserve to be mentioned here. Your help and advice have always been welcomed.

To all my friends, especially the ones sharing my passions: Fridays at Andrews/Arcís; never-ending nights killing minions in search of legendary items at Diablo 3; hiking, mountain biking, running and skiing (indeed I practice some sport); card games such as Magic the Gathering or Legend of the Five Rings; and many more. Thank you, I keep all of you in my heart.

I also want to acknowledge my whole family, for believing in me and encouraging me during the doctorate. And to those that are gone, I want to express my gratitude by these words. I'm sure that you'd also be proud of me right now. You have made this possible.

And at the end, thank you reader. If you are reading this line after the others, you at least read one page of this thesis.

Abstract

Protein-DNA interactions are indispensable players in the daily activities of cells. DNA-binding proteins regulate gene expression and are responsible of DNA replication, packaging, repair and recombination. Among them, transcription factors activate/repress gene transcription by binding to specific genomic sites. Hence, the characterization of transcription factor binding sites turns out to be crucial in order to understand gene regulation. In this context, the development of computational tools is foremost. Here, I show the prediction of redundant transcription factors in yeast using a combination of homology-based tools and protein-protein interactions. The approach was automated and incorporated into ModLink+, an online and user-friendly tool to infer the fold of remote homologs. Moreover, I describe split-statistical potentials for protein-DNA interactions. Finally, I present SHAITAN, a statistical/homology-based approach that can be used to both predict transcription factor binding sites and infer the more likely transcription factors to bind a DNA sequence of interest.

Resum

Les interaccions proteïna-ADN són indispensables en l’activitat diària de les cèl·lules. Les proteïnes que participen en aquestes interaccions s’encarreguen de la regulació de l'expressió gènica i són responsables de la replicació, l'empaquetament, la reparació i la recombinació de l’ADN. Entre aquestes proteïnes, els factors de transcripció activen/reprimeixen la transcripció de gens mitjançant la unió a llocs específics dins el genoma. Per tant, la caracterització dels llocs d'unió dels diferents factors de transcripció és crucial per tal d’entendre com funciona la regulació gènica. En aquest context, desenvolupar eines computacionals és importantíssim. En aquesta tesi predict redundància entre factors de transcripció de llevat eines fent servir eines basades en homologia i interaccions proteïna-proteïna. Aquesta aproximació va ser automatitzada i incorporada a

ModLink+, una eina accessible des d’internet i fàcil d'usar per a inferir el plegament de proteïnes a partir d’homòlegs remots. D'altra banda, descric potencials estadístics fraccionats per a interaccions proteïna-ADN. Finalment presento SHAITAN, una aproximació basada en homologia i potencials estadistics que pot ser utilitzada per a predir els llocs d'unió de factors de transcripció així com per saber quins factors de transcripció són més probables que s’uneixin a una determinada seqüència d'ADN.

Preface

This thesis represents the culmination of my work and learning during my doctorate. The journey began in form of collaboration with Dr. Anthony Gitter and Prof. Ziv Bar-Joseph from the Carnegie Mellon University (USA). In principle, I only had to apply the main concept behind my MSc project, which consisted in the use of protein-protein interactions for improving fold recognition in twilight zone proteins, to explain redundancy in gene regulatory networks. From that moment on, transcription factors and their interactions with DNA, rather than the prediction of protein fold or protein-protein interactions, became the center of my universe.

Although I would like to say that I walked my path through this thesis as straight as an arrow, I have to admit that this was not the case. All along, I have been given the opportunity to flirt with other interesting fields in systems biology such as network medicine or mutual information. The first project aimed to predict aneurysm candidates by integrating multiple sources of information, including protein-protein interactions and microarrays. In the second, after a crazy idea that came to my mind while attending to a seminar of a visiting fellow from Buenos Aires, Dr. Elin Teppa, I expected to improve fold prediction by comparing networks of mutual information derived from Pfam alignments. Both projects taught me a very important lesson: some things are just never meant to be no matter how much we wish they were[1].

By then, an ex-colleague of mine, Dr. Elisenda Feliu, was working on the application of statistical potentials to solve protein-protein docking, and my supervisor, Prof. Baldo Oliva, encouraged me to study protein-DNA interactions using the exact same approach. It was already 2012. I put a lot of effort and enthusiasm in this new project. To derive statistical potentials, the first thing one needs is a set of non-redundant structures. I did some research on the field and

[1] This is actually a good motto when it comes to women.

found a promising starting point: the set of structures from the work of Dr. Mu Gao and Prof. Jeffrey Skolnick. Nevertheless, the set was composed of specific and non-specific protein-DNA interactions, so the first time I tested my statistical potentials to predict the binding sites on a handful of transcription factors it resulted in a failure of biblical proportions. And what was worse, I had a poster to present to a congress in a few weeks!

It was during that congress, the RECOMB 2012, that I met Dr. Matt Weirauch. He provided me access to a database that contained hundreds of transcription factors and their binding sites. That was a point of inflection in my doctorate. The next week, my supervisor and I decided to incorporate such data into the statistical potentials and soon I started to obtain promising results. Since then, it has been one and a half years of trial and error until last week[2].

All in all, I am happy to have worked on such an interesting topic and I am glad that I have been given the opportunity to report my findings in this thesis. I truly hope you enjoy the reading!

[2] This is currently being written at 1:13AM on the 31st of August.

Table of contents

List of figures

List of tables

List of publications

Publications are listed in reverse chronological order. Articles 1, 2, 4 and 5 conform the "Results" section of this thesis. The remaining articles are mentioned, if necessary, along the different sections.

1. **Fornes, O.**, Weirauch, M.T., Hugues, T.R., & Oliva, B. **SHAITAN: On the prediction of protein-DNA interactions with a homology-based approach** (*To be submitted*).

2. **Fornes, O.**, Garcia-Garcia, J., Bonet, J., & Oliva, B. (2014). **On the Use of Knowledge-Based Potentials for the Evaluation of Models of Protein-Protein, Protein-DNA, and Protein-RNA Interactions.** *Advances in Protein Chemistry and Structural Biology,* 94, 77-120.

3. Garcia-Garcia, J., Bonet, J., Guney, E., **Fornes, O.**, Planas-Iglesias, J., & Oliva, B. (2012). **Networks of Protein-Protein Interactions: From Uncertainty to Molecular Details.** *Molecular Informatics*, 31(5), 342-362.

4. **Fornes, O.**, Aragues, R., Espadaler, J., Marti-Renom, M.A., Sali, A., & Oliva, B. (2009). **ModLink+: improving fold recognition by using protein-protein interactions.** *Bioinformatics,* 25(12), 1506-1512.

5. Gitter, A., Siegfried, Z., Klutstein, M., **Fornes, O.**, Oliva, B., Simon, I., & Bar-Joseph, Z. (2009). **Backup in gene regulatory networks explains differences between binding and knockout results.** *Molecular Systems Biology*, 5, 276.

6. Hofmann-Apitius, M., Fluck, J., Furlong, L., **Fornes, O.**, Kolárik, C., Hanser, S., Boeker, M., Schulz, S., Sanz, F., Klinger, R., Mevissen, T., Gattermayer, T., Oliva, B., & Friedrich, C.M. (2008). **Knowledge environments representing molecular entities for the virtual physiological human.** *Philosophical Transactions of the Royal Society A: Mathemathical, Physical and Engineering Sciences*, 366(1878), 3091-3110.

List of posters

1. Daniel Poglayen, Ana Garcia, Jascha Casadio, **Oriol Fornes**, Javier Garcia-Garcia, Guy Zinman, Manuel Alejandro Marin Lopez, Ziv Bar-Joseph, Heribert Hirt, Judith Klein-Seetharaman and Baldo Oliva. **High-throughput information integrated with in-silico predictions to identify key participants in host-pathogen interactions.** *XIII Spanish Symposium on Bioinformatics (JBI2014).* Sevilla, Spain, 2014.

2. **Oriol Fornes** and Baldo Oliva. **Prediction of DNA-binding specificities using statistical potentials.** *16th Annual International Conference on Research in Computational Molecular Biology, RECOMB2012*. Barcelona, Catalonia, 2012.

3. Jascha Casadio, **Oriol Fornes**, Elena Hidalgo, José Ayté, Isabel Calvo Arnedo, Patricia Garcia and Baldo Oliva. **A computational analysis of the regulation of oxidative stress genes in *S. pombe* by Pap1 and Prr1.** *16th Annual International Conference on Research in Computational Molecular Biology, RECOMB2012*. Barcelona, Catalonia, 2012.

4. David Alarcon, Aggeliki Kosmopoulou, Jaume Bonet, **Oriol Fornes**, Roberto Mosca, Jose Manuel Mas, Patrick Aloy and Baldo Oliva. **Exploring type-2 Diabetes Protein Interaction Networks by modeling their complex structures.** *17th Annual International Conference on Intelligent Systems for Molecular Biology and 8th European Conference on Computational Biology, ISMB/ECCB09*. Stockholm, Sweden, 2009.

5. **Oriol Fornes**, Ramon Aragues, Jordi Espadaler, Jaume Bonet, Marc A. Marti-Renom, Andrej Sali and Baldo Oliva. **ModLink+: improving fold recognition by using protein-protein interactions.** *3DSIG 2009: The 5th Structural Bioinformatics and Computational Biophysics, ISMB satellite meeting*. Stockholm, Sweden, 2009.

1. Introduction

“Read and find out.”
-Robert Jordan

In the midst of the Second World War, it was shown for the first time that deoxyribonucleic acid (DNA) is the material of inheritance (1). Until then, biologists thought that genes, the units of inheritance, were made of protein. But it wasn’t until 1953, an *annus mirabilis* for science, that the structure of DNA was disclosed. James Watson and Francis Crick were the first to describe the double-helical structure of DNA (2) and suggested it as a possible copying mechanism for the genetic material. In the same issue of *Nature*, Wilkins, Stokes and Wilson found evidence that the structure existed in biological systems (3), and Rosalind Franklin and Ray Gosling provided further evidence of the helical nature of nucleic acids and concluded that the phosphate backbone was placed on the outside of the structure (4). Later that year, Watson and Crick followed up with a largely accurate speculation on how base pairing in the double helix allowed DNA replication (5), and Franklin and Gosling detailed the differences between the A and B structures of DNA (6). Since then, many important biological discoveries, including some Nobel Prizes[3], have revealed that the function of DNA depends on its interaction with proteins.

[3] In 1959, Arthur Kornberg won the Nobel Prize in Medicine (shared with Severo Ochoa) for the discovery of the DNA polymerase. In 1965, François Jacob and Jacques Monod won the Nobel Prize in Medicine (shared with André Lwoff) for the discovery of the *lac* operon (see further in “Repressors” in section 1.1.1). In 1995, Christiane Nüsslein-Volhard and Eric F. Wieschaus won the Nobel Prize in Medicine (shared with Edward B. Lewis) for the discovery of a set of transcription factors crucial for fruit fly development. In 2001, Leland H. Hartwell, Tim Hunt and Sir Paul M. Nurse won the Nobel Prize in Medicine for the discovery of the main CDK/cyclin that regulate the cell cycle. In 2009, Carol W. Greider and Elizabeth H. Blackburn won the Nobel Prize in Medicine (shared with Jack W. Szostak) for the discovery of the enzyme that protects the telomeres in chromosomes (see further in “Termination” in section 1.1.2).

1.1 Biological relevance of protein-DNA interactions

Protein-DNA interactions (PDIs) play an essential role in the daily activities of cells. PDIs are involved in the regulation of gene expression and participate in the replication, packaging, repair and recombination of DNA. PDIs can be either specific or non-specific, depending on whether the protein recognizes a particular DNA sequence or not. Among DNA-binding proteins, transcription factors (TFs) are the most widely studied. By binding to specific DNA sequences, they can either promote or repress gene transcription. Some enzymes can also bind to DNA, and among them, the polymerases that copy DNA along transcription and replication are of particular importance. This section revises the main processes involving PDIs.

1.1.1 Regulation of gene expression

The regulation of gene expression allows living organisms to express proteins when it is required. It includes a wide range of mechanisms that control the production of ribonucleic acid (RNA) and proteins. Almost any step of gene expression can be regulated, from transcriptional initiation to post-translational modification of proteins, passing through the processing of RNA. However, for the purpose of this thesis, I only focus on transcriptional regulation by PDIs. Proteins responsible for such regulation are known as regulatory proteins. They usually bind to DNA sites located near the promoter region of genes, although it may not always be the case (see "Enhancers" and "Silencers"). Moreover, by interacting with their binding sites, these proteins are able to affect gene transcription by RNA polymerase. As the main participants of gene expression, TFs will be revised in next section (refer to section 1.2). There are different mechanisms by which regulatory proteins control the transcription of genes (see below).

Specificity factors

In prokaryotes, sigma factors have the ability to alter the specificity of RNA polymerase for a given promoter (7). The use of a specific sigma factor to initiate the transcription of a gene depends on the gene and on the environmental signals needed to initiate the transcription of that gene. For example, in *E. coli*, when the bacteria are subjected to heat stress, the σ^{32} protein suffers a conformational change that causes RNA polymerase to bind to a set of specialized promoters that regulate genes coding for proteins involved in heat-shock response (8).

Activators

TFs that promote the expression of a gene or a set of genes are known as activators. They usually bind to sequence-specific DNA sites located near the promoter, thereby facilitating the binding of the transcription machinery. In order to increase the expression of their regulated genes, activators can either interact with RNA polymerase subunits or distort the structure of DNA (see further in "The TATA-box promoter" in section 1.2.1).

Enhancers

Activators can bind to DNA sites located up to 1 megabase away from the promoter known as enhancers (9), or even in another chromosome (10), in order to loop DNA and promote gene transcription. Enhancers are very important during development and are involved in many different processes such as segmentation in invertebrates (11), or the establishment of the body axes in vertebrates (12–14). It is thought that human cells may contain up to 1 million active enhancers (15). Sometimes, multiple enhancers associate forming large clusters in order to define cell identity. For example, in mouse embryonic stem cells, Oct4, Sox2 and Nanog, along with other TFs, have been associated to 231 different genes most of which control the pluripotent state (16). These clusters of enhancers are known as "super-enhancers" and have been also found at oncogenes in cancerous cells (17).

Repressors

As their name suggests, these TFs repress gene expression by binding to non-coding DNA sequences that are close to or over the promoter region, in order to block the progress of RNA polymerase upon transcription. For example, in *E. coli*, the binding of the *lac* repressor protein in the major groove of the promoter region of the *lac* operon prevents the binding of RNA polymerase, which blocks the synthesis of enzymes that digest lactose when there is no lactose available in the environment (18). In the presence of lactose, the bacteria generate allolactose, which then binds to the repressor and causes it to detach from DNA.

Silencers

As for activators, repressors can bind to DNA sites located far away from the promoter known as silencers. Silencers work in a similar way than enhancers but, instead, they silence gene expression (19). For example, in T lymphocytes, a silencer is responsible for lineage-specific differential expression of CD4 during development, which results in either helper or cytotoxic T cells (20).

General transcription factors

Also known as basal TFs because they are always present, the function of these proteins is to position RNA polymerase at the start site of a protein-coding sequence and then release it to transcribe the mRNA (21). For example, in eukaryotes, RNA polymerase II requires the binding of 6 different general TFs in order to initiate gene transcription (22): TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH (see further in "The TATA-box promoter" in section 1.2.1).

Chromatin structure

In eukaryotes, a particular case of regulation is mediated by the accessibility of DNA to RNA polymerase. The level of packaging of the chromatin (see section 1.1.3) dictates which DNA regions can be transcribed and which cannot (23). Chromatin can be unpacked as a result of histone modifications, including acetylation, methylation, phosphorylation or ubiquitination. Therefore, cells can up- or down-regulate the expression of genes by packing/unpacking DNA (24).

1.1.2 DNA replication

The replication of DNA is the basis for biological inheritance and it occurs in all living organisms. It is a process by which cells generate two identical replicas from one original DNA molecule. Each strand of the original DNA molecule is used by a DNA polymerase as template for the production of a complementary strand. Therefore, each produced replica contains one strand from the original DNA molecule. Overall, the mechanism of replication is very precise, with DNA polymerases making less than one mistake for every 10^7 nucleotides added (25). Moreover, some DNA polymerases also have the ability to remove nucleotides from the end of the growing strand in order to correct mismatches. Other post-replication mismatch repair mechanisms also monitor DNA for errors, and are able to identify mismatches in the newly synthesized DNA strand from the original strand sequence. By combining these three factors, cells achieve a replication fidelity of less than one mistake for every 10^9 nucleotides added (25).

Initiation

DNA replication is initiated at particular points in DNA known as “origins”, which are bound by initiator proteins (26). For example, in *E. coli*, this protein is DnaA, while in yeast it requires the origin recognition complex (27). These proteins recognize DNA motifs that tend to be “AT-rich” since A-T pairs only have two hydrogen bonds (rather than the three formed in a C-G pair), which make them easier to separate. The binding of initiator proteins to the origin is required in order to recruit other proteins and form the pre-replication complex (28), which unzips the double-stranded DNA.

Elongation

Once the two strands of DNA are separated by helicases, a primase adds a short fragment of RNA, named primer, to each template strand. During the replication process, the leading strand receives only one RNA primer, while the lagging strand receives several. This is because the leading strand is extended continuously by a high processivity, replicative DNA polymerase, while the lagging strand is extended discontinuously from each primer as the double-stranded DNA unzips, generating short DNA fragments, typically between 100-200 nucleotides long, known as Okazaki fragments (29).

After DNA extension, an RNase removes all primers and a low processivity DNA polymerase, distinct from the one used for the replication, fills the gaps left. In the end, the process leaves a single nick on the leading strand and several nicks on the lagging strand, which are filled by a DNA ligase, completing the replication. There are other proteins involved in elongation, such as topoisomerases or gyrases, which relax the unzipped DNA from helicases, or DNA clamps, which form a sliding clamp around DNA and prevent the polymerase from dissociating from the template strand.

Termination

In eukaryotes, the process of replication is unable to reach the very end of chromosomes. Instead, it stops at the telomere region of the chromosome, which contains repetitive DNA, shortening the telomere of the replicated DNA strand. This is a normal process in somatic cells and, as a result, cells can only divide a certain number of times before the loss of DNA prevents further division (30). However, in germ cells, since they are in charge of passing the genetic material to the next generation, a telomerase extends the repetitive telomeric regions to prevent its degradation (31).

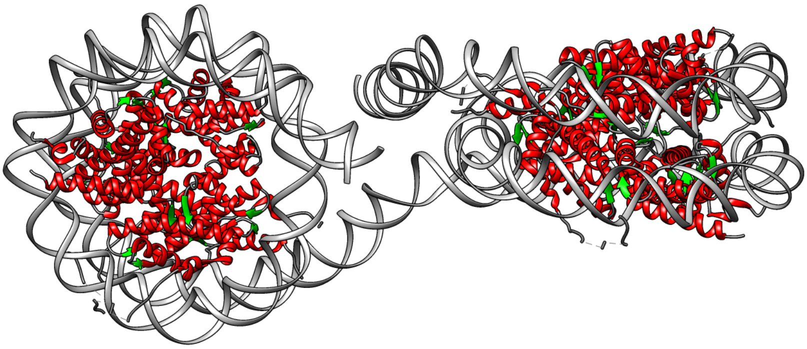
1.1.3 DNA packaging

In order to understand the level of packaging that DNA suffers in the nucleus of cells, lets first put the size of DNA into context. Each human somatic cell contains approximately 6 billion base pairs of DNA. Moreover, a base pair is around 0.34 nanometers long, which makes a total of 2 meters of DNA per cell. Furthermore, the human body is estimated to contain about 50 trillion cells, which means that the amount of DNA stored in each human being arises to the astronomic number of 100 trillion meters. Now, considering the fact that the Sun is 150 billion meters from Earth, human cells altogether contain enough DNA to go to the Sun and back more than 300 times, or to circle the Earth’s equator 2.5 million times. This is possible because DNA is highly compacted and packed in the nucleus of each cell in form of chromosomes.

An initial step of condensation is due to nucleosomes. They are composed of two of each of the histones H2A, H2B, H3, and H4, which bind and wrap about 1.67 turns of DNA, or 146 base pairs (32). The addition of the linker histone H1 to the base of the nucleosome wraps an additional 20 base pairs, which results in two full turns of DNA around the nucleosome, or 166 base pairs (33). The packaging of DNA into nucleosomes shortens the fiber of DNA by seven fold. In other words, the length of DNA per cell is reduced from 2 meters to 28 centimeters. However, this is still too much to be fit in the nucleus of a cell. Another step of condensation is achieved when the nucleosomes fold into a fiber of approximately 30 nanometers in diameter (34), which in turn form loops of 300 nanometers length in average (35). These 300 nanometer fibers are further compressed and folded to produce a 250 nanometer-wide fiber, which is tightly coiled into the chromatid of the chromosome. However, DNA is only condensed at that level during cell division, especially in the metaphase

Some processes such as transcription or replication also require the chromosomes to uncoil so that the two strands of DNA can come apart temporarily, thus allowing polymerases access to the DNA template. Nevertheless, the presence of nucleosomes and the folding of chromatin into 30-nanometer fibers suppose a physical barrier for the enzymes that unwind and copy DNA. Therefore, there are two major mechanisms that open up chromatin fibers and/or remove histones transiently, which make the chromatin more accessible for DNA transcription and replication. On the one hand, histones can be enzymatically modified by the addition of acetyl, methyl, or phosphate groups, which releases them from DNA (36) and, on the other hand, histones can be displaced by chromatin remodeling complexes (37).

Figure 1.1. Structural view of two concatenated nucleosomes (38).



The structure shows that linker DNA zigzags back and forth between two stacks of nucleosome cores. By successively stacking to one on another, nucleosomes form continuous 30 nanometer fibers.

1.1.4 DNA repair

Every day cells must face tens of thousands of DNA lesions. Although environmental causes such as chemical agents, ultraviolet (UV) radiation, or ionizing radiation are the main sources of DNA damage, normal metabolic processes inside the cell as well as spontaneous depurination of DNA also play an important role (39). All in all, human cells have to repair about 0.000165% of the 6 billion base pairs of their genome daily. However, although it may seem a small percentage, any unrepaired lesions in critical genes, such as tumor suppressor genes, can impede the cell to function normally, or increase the likelihood of tumor formation and contribute to its heterogeneity.

Most DNA damage affects the primary structure of the double helix by chemical modification of the bases. These modifications can disrupt the structure of DNA by introducing non-native chemical bonds or bulky adducts that do not fit in the double helix standards. Moreover, when the damage is near nucleosomes, DNA packaging can result affected. The typical DNA damage due to endogenous cellular processes comprises oxidation, alkylation, hydrolysis, formation of adducts, or mismatches during replication. In contrast, environmental damage of DNA comes in many forms such as UV light (creates pyrimidine dimers), ionizing radiation (causes irreversible DNA breaks), or industrial chemicals. Obviously, cells cannot function if any damage affects the integrity or accessibility of the essential information that is contained in their genomes. Therefore, there exist several mechanisms to repair DNA (40–47).

1.1.5 DNA recombination

Genetic recombination is a process by which two chromosomes, or two different regions from the same chromosome, exchange genetic material. In sexually reproducing organisms, it is essential in order to ensure genetic variation within a population (48). It also allows lymphocyte B cells to change the class of an antibody (49), and it is one of the mechanisms by which cells repair double-breaks in DNA molecules (45,46). The basic steps of recombination involve the alignment of two homologous DNA strands, the precise breakage of each strand, the exchange of genetic material between the strands, and the sealing of the resulting recombined molecules. Although it is very complex, this process occurs frequently in both prokaryotic and eukaryotic cells, and with a high degree of accuracy.

1.2 Transcription factors

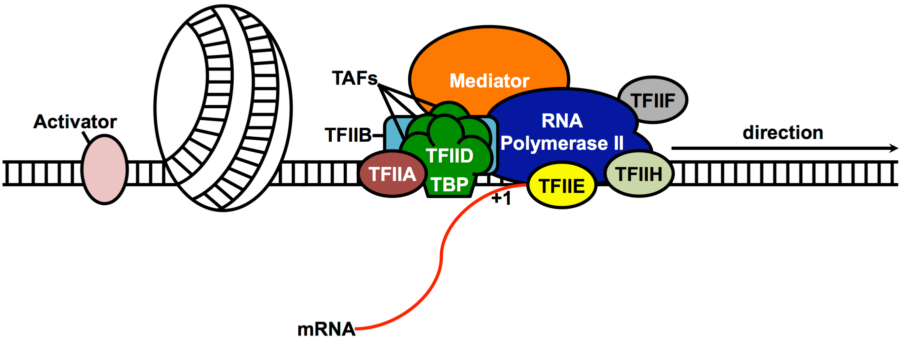
Among DNA-binding proteins, TFs are the *crème de la crème*. They are essential for the transcription of genes and therefore, they are present in all living organisms. In fact, due to their importance, mutations affecting TFs have been directly associated to diseases and cancer (50). For example, one third of human developmental disorders have been associated to dysfunctional TFs (51). The number of TFs found across species, as well as their proportion with respect to the number of genes, grow with their genome size (52). In eukaryotes, they account for approximately 5-10% of genes (53–57), and the human genome alone is estimated to encode around 2,000 TFs (58). TFs are able to bind alone or cooperatively to enhancer or promoter DNA regions adjacent to their regulated genes in order to promote/repress gene transcription. For example, TFs can stabilize/block the binding of RNA polymerase to DNA, promote the acetylation/deacetylation of histones to pack/unpack DNA, or recruit coactivators/corepressors to the TF-DNA complex (see "Activators" and "Repressors" in section 1.1.1). In eukaryotes, TF-binding sites range between 6 and 10 base pairs and are usually degenerated.

Regarding gene transcription, in eukaryotes it begins with the binding of a TF to its cognate site. This is followed by sequential recruitment of general TFs and ultimately, of RNA polymerase II. Altogether, these proteins compose the transcription preinitiation complex (PIC). In prokaryotes, RNA polymerase is able to bind the promoter of a gene *per se*. Next steps include promoter melting and escape, transcript elongation until termination sites, and additional capping and processing of the nascent transcript, which occurs co-transcriptionally. For the purpose of this thesis, this section only revises gene transcription until the ensemble of the PIC. A typical example to illustrate the PIC formation upon transcription is the recruitment of general TFs by the TATA-box binding protein (TBP) (see Figure 1.2).

1.2.1 The TATA-box promoter

The TATAAA sequence present in the TATA-box promoter is recognized by TBP, which binds to the promoter and distorts the structure of DNA, thus facilitating the recruitment of several TBP-associated factors (TAFs) (see Figure 1.3). The recruitment of TBP is regulated both positively and negatively. On the one hand, activators increase the binding of TBP to the TATA-box during transcriptional activation (59). On the other hand, negative factors such as Mot1 or the Taf1 N-terminal domain suppress the DNA-binding activity of TBP (60).

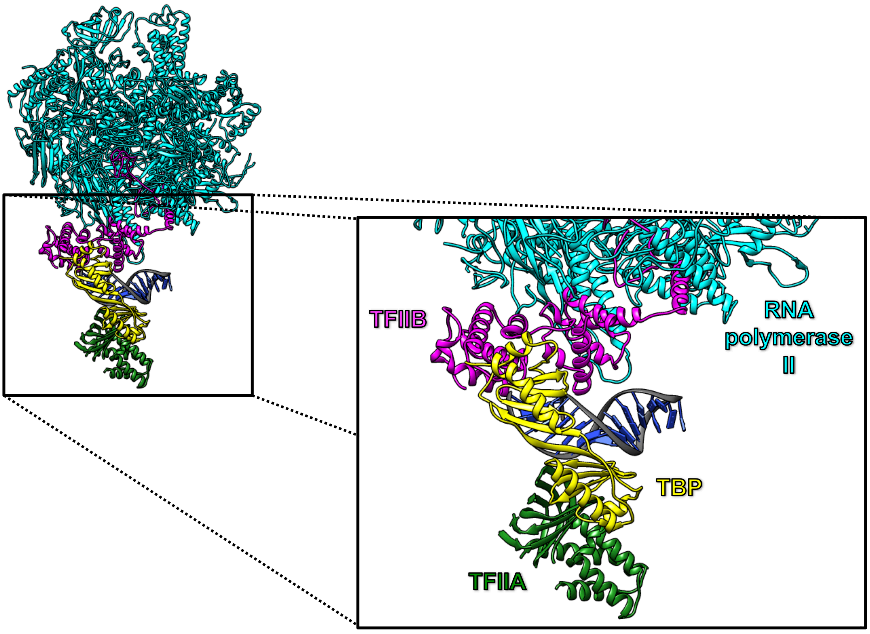
Figure 1.2. Illustration of a TATA-box promoter.



Apart from TAFs, general TFs associate with TBP at the TATA-box. Initially, the binding of TFIIA stabilizes the interaction between TBP and DNA. In yeast, TAF40 interacts with TFIIA and adds TFIID to the complex (61). Moreover, TFIIA competes with negative factors such as NC2, Mot1 and Taf1 for binding to TBP. Next, TFIIB, binds to the flanking regions immediate to the TATA-box, which is critical for the formation of a stable PIC (62). A crystal structure at 4.5 Å of RNA polymerase II complexed with the N-terminal region of TFIIB revealed a loop called the “B-finger” that reaches into the catalytic center of the polymerase, where it interacts with both DNA and the nascent RNA (63). More recent crystals have shown that the C-terminal region of TFIIB is located above the polymerase active center cleft, which directs the TFIIB

N-terminus towards the catalytic center. A linker helix/strand of TFIIB interacts with the polymerase rudder and assists in opening DNA at the active center. Furthermore, at the catalytic center of the polymerase, the DNA template strand slides into the cleft and is scanned for the transcription start site with the help of a helical region of TFIIB, the "B-reader", which collaborates in the start site selection. As the nascent transcript grows, at the length of 5 nucleotides, it contacts the B-finger to form a stable complex. In addition, at the length of 7 residues, the nascent transcript clashes with the B-finger and displaces TFIIB, which leads to promoter escape (64,65). The PIC is further stabilized with the binding of RNA polymerase II and TFIIF. Finally, the general TFs TFIIH and TFIIE are recruited to the PIC together with the mediator complex, and transcription initiates. TFIIH contains 10 subunits that control an ATP-dependent transition from the closed to open PIC, which is required for productive transcription initiation (66).

Figure 1.3. Structural view of a TATA-box promoter.



TATA-box complex with TBP (67) (yellow), TFIIB (68) (magenta), TFIIA (69) (green) and RNA polymerase II (65) (cyan), created by structural superimposition with the UCSF Chimera package (70).

However, the majority of eukaryotic promoters lack a canonical TATA-box. In yeast, only about 20% of genes contain a TATA-box, and most of them are associated to stress response. In contrast, TATA-less promoters contain other elements such as Initiator and the downstream promoter element that are used for promoter recognition by the transcription machinery (71). The majority of TATA-less promoters are found in basic housekeeping genes (72).

1.2.2 Structure of transcription factors

TFs are modular, and are constituted of 3 different domains. The DNA-binding domain (DBD) allows the TF to attach to DNA through a combination of electrostatic (of which hydrogen bonds are a special case) and Van der Waals forces. This allows the TF to recognize DNA in a sequence specific manner. However, not all bases in the binding site may actually interact with the TF, and some of these interactions may be weaker than others. Thus, a TF is capable of binding a subset of closely related sequences, each with different affinity, which allows to represent its binding sites as a probabilistic model named position weight matrix (PWM). For example, although the consensus site for TBP is TATAAA, it also binds to similar sequences like TATATA or TATAAT (73). The limited number of DBDs that exist in nature has been used to classify TFs in different families (see Table 1.1 and Figure 1.4).

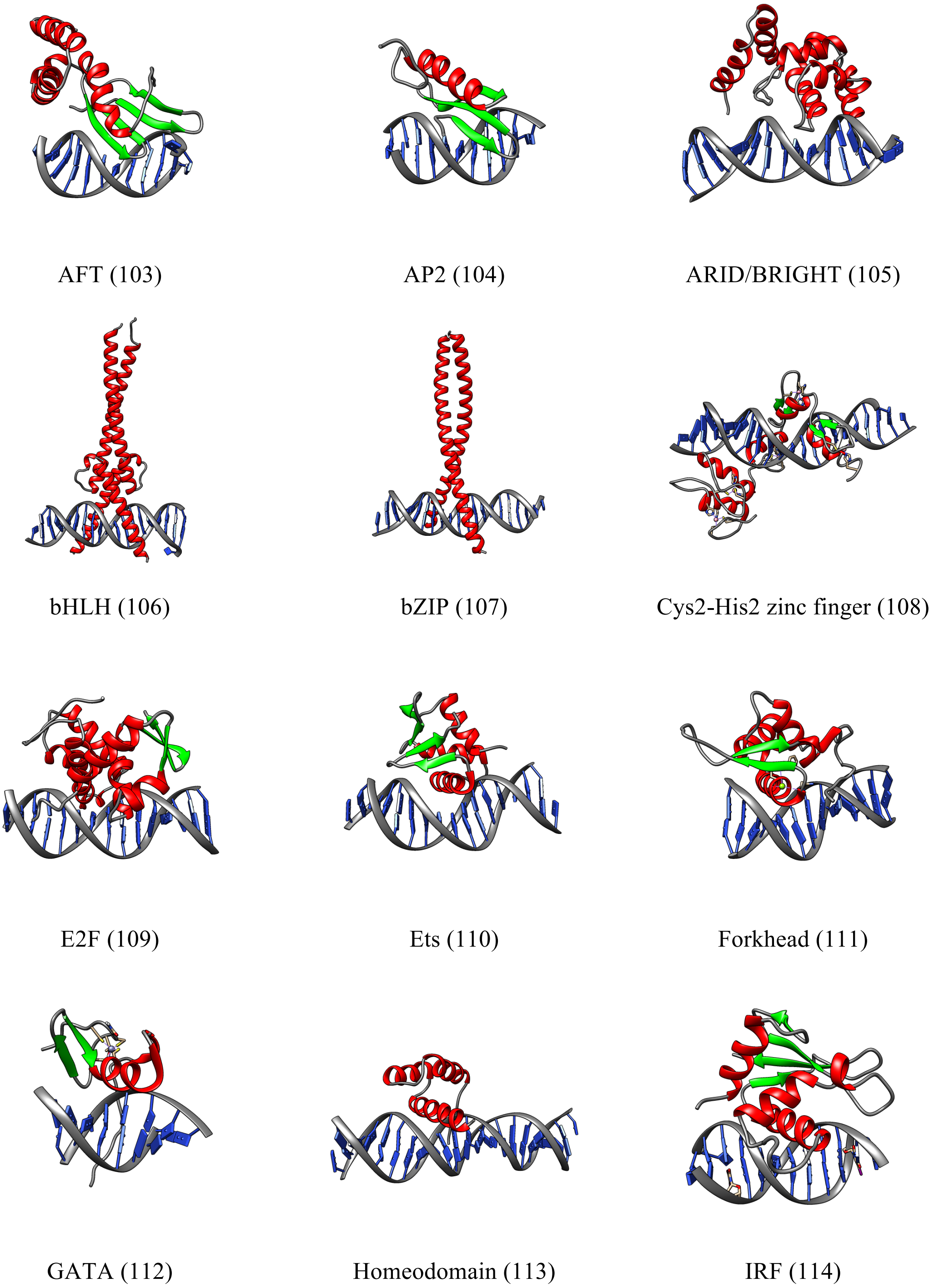
The trans-activating domain allows the TF to interact with other transcriptional coactivators such as TAF9, MED15, CBP and p300. For example, the nine-amino-acids trans-activation domain, which appears in a large superfamily of eukaryotic TFs represented by Gal4, Oaf1, Leu3, Rtg3, Pho4, Gln3 and Gcn4 in yeast, and by p53, NFAT, NF-κB and VP16 in mammals, interacts directly with the general coactivators TAF9 and CBP/p300 (74).

The signal-sensing domain, which is not present in all TFs, senses external signals and transmits them to the rest of the transcription

Table 1.1. Structural classification of TFs by their DNA-binding domains (only the most relevant for this thesis are shown).

Family	**Function**
AFT	iron homeostasis (75)
AP2	early development, apoptosis, cell cycle (76)
ARID/BRIGHT	cell growth, cell differentiation, development (77)
bHLH	neurogenesis, cardiogenesis, hematopoiesis, cell cycle, myogenesis (78)
bZIP	cell differentiation (79), spermatogenesis (80), stress response (81), steroidogenesis (82)
Cys2-His2 zinc finger	transcription regulation, RNA binding (83)
E2F	cell cycle (84)
Ets	cell cycle, apoptosis, cell differentiation (85)
Forkhead	cranio-pharingeal, hair, ear, speech and language development (86)
GATA	hematopoiesis, cardiac development (87)
Homeodomain	patterning, cell differentiation, development (88)
IRF	immunity, oncogenesis (89)
MADS box	muscle development, cell differentiation, arginine metabolism, flowering (90)
Myb/SANT	cell proliferation and differentiation (91)
NAC/NAM	plant development (92)
Nuclear receptor	development, homeostasis, reproduction, metabolism (93)
Paired box	kidney, eye, ear, nose, limb muscles, vertebral column and brain development (94)
POU	nervous system development (95), neuroendocrine system (96)
SMAD	signal transducers (97)
Sox	sex differentiation, neural development (98)
T-box	tissue differentiation, development (99)
TBP	stress response (100)
WRKY	plant immunity and defense (101)
Zinc cluster	DNA recognition, RNA packaging, transcriptional activation, apoptosis, protein folding/assembly, lipid binding (102)

Figure 1.4. Structural view of a member of each family of TFs from Table 1.



AFT (103)

AP2 (104)

ARID/BRIGHT (105)

bHLH (106)

bZIP (107)

Cys2-His2 zinc finger (108)

E2F (109)

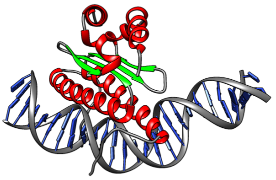
Ets (110)

Forkhead (111)

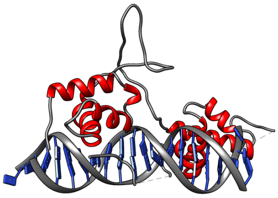
GATA (112)

Homeodomain (113)

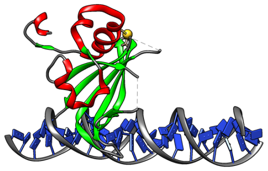
IRF (114)



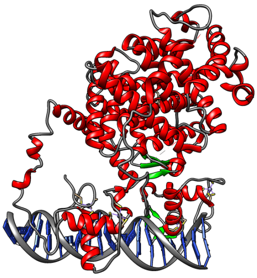
MADS box (115)



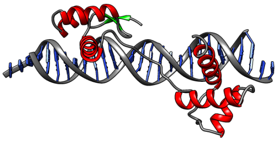
Myb/SANT (116)



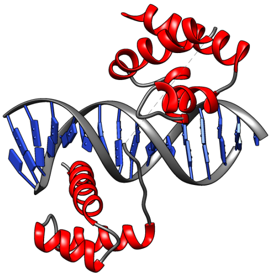
NAC/NAM (117)



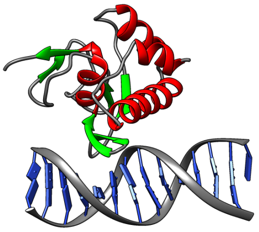
Nuclear receptor (118)



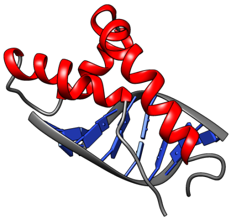
Paired box (119)



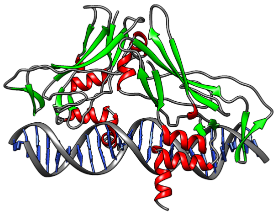
POU (120)



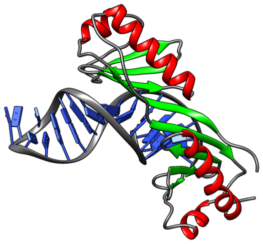
SMAD (121)



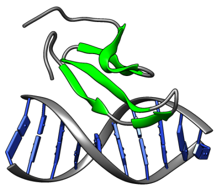
Sox (122)



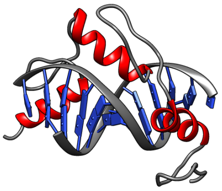
T-box (123)



TBP (67)



WRKY (124)



Zinc cluster (125)

complex, which results in up- or down-regulation of the expression of the transcribed gene. For example, a PAS domain allows the hypoxia-inducible factor 1 to mediate the transcriptional activation of the erythropoietin gene in front of decreased O_2 levels (126).

1.2.3 Protein interactions of transcription factors

Protein-protein interactions (PPIs) are essential in order to regulate gene expression. Not only they are necessary upon PIC formation, but also for the transduction of external signals into the expression of one gene or another. Many TFs involved in developmental processes, such as Smad (97), STAT (127), β-catenin (128) or NF-κB (127), are mainly signal transducers. Typically, these TFs are found in the cytoplasm and, upon activation, they translocate to the nucleus and promote the transcription of their regulated genes. As signal transducers, these TFs interact with different proteins, including membrane receptors, adaptor proteins, or kinases.

Some TFs bind to DNA as dimers, as it is the case of the bHLH (78), bZIP (129) and nuclear hormone receptor families (130). Dimerization increases the specificity and affinity of TFs for DNA and allows them to interact with different proteins (131). In some cases, different combinations of monomers can transform the dimer from one that activates gene transcription to one that represses it. For example, the CCAAT/enhancer binding protein (C/EBP) is involved in various processes, such as adipogenesis, hematopoiesis and regulation of cell cycle. C/EBP belongs to the bZIP family and interacts with DNA in form of homo- or heterodimer. However, when it dimerizes with the C/EBP homologous protein 10, its DNA-binding activity is attenuated (132).

Finally, in eukaryotes, it is by now fairly clear that TFs do not act alone. Instead, they act cooperatively forming "enhanceosomes", which are assemblies of TFs stabilized by protein-protein as well as PDIs. A well known enhanceosome occurs in the human interferon-beta gene (133). Upon viral infection, NF-κB, the ATF-2/c-Jun dimer, and an interferon regulatory factor such as IRF-3/7 bind cooperatively to a nucleosome-free region of the gene promoter. A fourth protein, HMG-I, stabilizes the complex by promoting inter-protein interactions. Once assembled, the enhanceosome recruits the RNA-polymerase II machinery to the promoter and activates the transcription of the gene.

Characterization of protein-protein interactions

Several methods have been developed in order to identify PPIs. Protein complementation assays (PCA) (134) represent the biggest group of these methods. In a typical PCA protocol, the proteins of interest ("bait" and "prey") are covalently linked at the genetic level to incomplete fragments of a third protein known as the "reporter". Usually, the reporter protein is a TF that regulates a certain gene, which upon activation promotes an observable phenotype. The whole system is expressed *in vivo*. If the bait and prey proteins interact, the reporter fragments are close enough to become functional and, consequently, the reporter activity is detectable. Among PCA methods, the most widely used are the yeast two-hybrid assay (135) and tandem affinity purification (136).

Other methods are based in fluorescence and are aimed to detect weak and transient PPIs as occurring in living cells. Among them, the biomolecular fluorescence complementation (BiFC) (137) and the green fluorescence protein (GFP) (138) assays are the most popular. Förster/fluorescence resonance energy transfer (139) is another common flourescence-based assay in which the energy from a donor fluorophore is transferred to an acceptor fluorophore if they are in close proximity and properly oriented. A more sensitive method, the bioluminescence resonance energy transfer system (140), replaces the donor fluorophore by a luciferase.

A different family of methods is based on the array technology. The array contains several covalently attached proteins (probes), and their ability to interact with other labeled proteins (samples) is tested (141). A variation of these methods, the surface plasmon resonance system (142), does not require labeled samples. Instead, an optical biosensor identifies PPIs by detecting changes in the local refractive index, thus providing real-time affinity and kinetic data.

Some methods are able to characterize PPIs at atomic resolution, but they will be described in "High-resolution methods" in section 1.3.4. Table 1.2 summarizes the most commonly used methods to detect PPIs (this later group is not included).

Table 1.2. Most common experimental methods for detecting PPI.

Method	**Type**	**Binary**	**Complex**	**HT**
Yeast two-hybrid (135)	PCA	•		•
Mammalian PPI trap (143)	PCA	•		
Tox-r dimerization assay (144)	PCA	•		
Bimolecular fluorescence complementation (137)	PCA, fluorescence	•		
Proximity ligation assay (145)	PCA	•		
Förster/fluorescence resonance energy transfer (139)	Fluorescence	•		
Bioluminescence resonance energy transfer (140)	Fluorescence	•	•	
Protein microarrays (141)	Array	•	•	•
Surface plasmon resonance (142)	Array	•	•	
Tandem affinity purification (136)	PCA	•	•	•

Columns 3-5 use a dot to indicate if the method can detect binary PPI (Binary), multiple protein complexes (Complex), or if it can be used in a high-throughput manner (HT).

Databases of protein-protein interactions

The emergence of high-throughput techniques has allowed PPI detection at large scale. All this data can be found in various curated depositories, including the database of interacting proteins (DIP) (146), the biomolecular interactions network database (BIND) (147), the biological general repository for interaction datasets (BioGRID) (148), the human protein reference database (HPRD) (149), or MIntAct (150) (this subject is reviewed in (151); see Appendix 1).

1.3 Characterization of protein-DNA interactions

The characterization of PDIs is foremost in order to understand how proteins interact with DNA and regulate gene expression. This section is focused on different methods for the detection and characterization of PDIs, including the prediction of DNA-binding ability of proteins, the identification of protein-DNA association, and the elucidation of specific binding amino acids and nucleotides involved in the PDI. According to the nature of the experiment, these methods have been classified into 5 categories: *in vitro*, *in vivo*, high-throughput, high-resolution and computational.

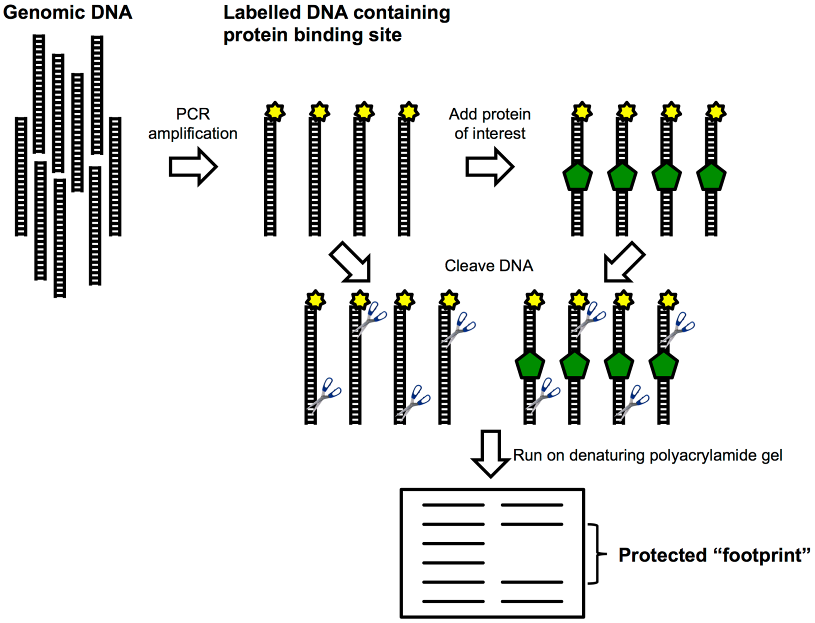
1.3.1 *In vitro* methods

Traditionally, *in vitro* methods have been employed to identify and characterize PDIs. Among them, the most widely used are DNA footprinting, southwestern blot, electrophoretic mobility shift assay and DNA pull-down.

DNA footprinting

This method relies on the principle that a protein bound to DNA will protect that DNA from enzymatic cleavage. The DNA region of interest is amplified and labelled using polymerase chain reaction (PCR) method (152). Then, a given protein is added to a portion of amplified DNA (the other portion is saved for later comparison) together with a cleavage agent (which is added to both portions). There are various cleavage agents that can be used such as DNase I (153), hydroxyl radicals (154) or UV radiation (155). If the protein does not bind to the target DNA, both samples, the one with protein and the one without protein, will show a ladder-like distribution when run on a polyacrylamide gel electrophoresis. On the contrary, if the protein binds the DNA, the protein sample will show a ladder distribution with a break in it, the “footprint”, where DNA has been protected from the cleavage agent. By varying the concentration of DNA-binding protein, the binding affinity of the protein can be estimated according to the minimum concentration of protein at which the footprint is observed.

Figure 1.5. Overview of DNA footprinting workflow.



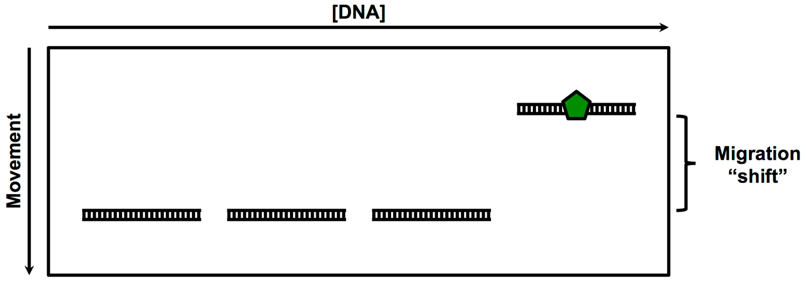
Southwestern blot

Soutwestern blotting involves a modified western blot procedure using labeled oligonucleotides instead of antibodies as probes (156). The whole cell content (raw or purified) is resolved on a denaturing polyacrylamide gel, which is followed by electrophoretic transfer of the proteins to a membrane, under renaturing conditions, and further incubation with the DNA sequence of interest. Then, the membrane is developed and photographed, revealing a band corresponding to the bound DNA sequence. The alignment of the band with the gel marks the position of the DNA-binding protein and provides information about its molecular weight. Additional western blotting or mass spectrometry can be used to identify the protein, if necessary.

Electrophoretic mobility shift assay

The electrophoretic mobility shift assay (157) is based on the observation that protein-DNA complexes migrate more slowly than free DNA molecules in non-denaturing polyacrylamide or agarose gel electrophoresis. Then, if a protein binds a given DNA molecule, a migration "shift" relative to the non-bound DNA probe can be observed. The specificity of the binding can be determined through a competition reaction between unlabeled and labeled different DNA sequences, which results in a decrease in the shifted signal if the different DNAs compete for the binding of the same protein. Alternatively, the protein-DNA complex can be crosslinked and the reaction run on a denaturing gel, where the specificity of the binding can be determined through visualization of a single shifted band. Moreover, if the starting concentrations of both the protein and DNA are known, as well as the stoichiometry of the complex (*i.e.* if the protein binds as a monomer or as a multimer), the apparent affinity of the protein for the DNA sequence can also be determined. Additionally, calculating the apparent affinities of different mutants and comparing them with respect to the wild type can also be used to identify the key DNA-binding amino acids of the protein.

Figure 1.6. Illustration of EMSA principle.



DNA pull-down

In a DNA pull-down assay, double-stranded DNA sequences are labeled with high affinity tags, such as biotin, in order to immobilize them in functionalized beads or columns (usually containing avidin[4] or derivates). The biotin-tagged DNAs are then incubated with proteins from a cell lysate. The formed protein-DNA complexes are further purified using agarose or magnetic beads. Once the proteins are eluted from DNA, they can be detected by western blotting or identified by mass spectrometry. Alternatively, the protein can be labeled with an affinity tag or the protein-DNA complex can be isolated using an antibody against the protein of interest. In this case, the unknown DNA sequence bound by the protein can be detected by southern blot or through PCR analysis.

1.3.2 *In vivo* methods

In contrast to the previous techniques, some methods allow the characterization of PDIs as occurring in living cells. Among them, chromatin immunoprecipitation (ChIP) is the most popular (158), but I will describe it together with other high-throughput methods (see section 1.3.3).

Proximity ligation assay

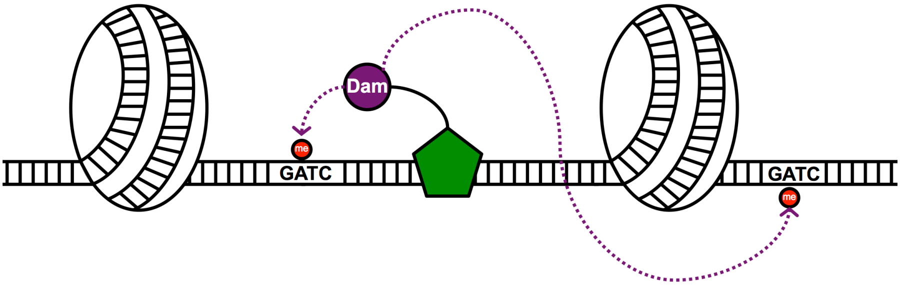
Proximity ligation assay (PLA) is an ultrasensitive technique for measuring any kind of interaction *in vivo* (145). For PDIs, it requires two primary antibodies grown in different species, each of them specific for either the protein of interest or DNA (the second antibody is unspecific for any DNA sequence), which are then bound by specie-specific secondary antibodies named PLA probes. Each PLA probe contains a unique short DNA sequence attached to it and, if the two probes are in close proximity, the addition of two other DNA fragments forms a circle. Once ligated, the circular DNA is amplified using labeled nucleotides, resulting in a bright spot when observing the cells or tissue sample with a fluorescence microscope.

[4] The avidin-biotin complex is the strongest known non-covalent interaction between a protein and a ligand ($K_d = 10^{-15}$ M), which makes it one of the most useful tags for the purification and detection of complexes.

DNA adenine methyltransferase identification

In this technique, a protein of interest is fused with a bacterial DNA adenine methylase (DAM) (159). Upon binding, the protein brings DAM to its DNA recognition sites, which results in the methylation of adenines in neighbouring GATC sites. Further digestion of the genome with the enzyme DpnI, which cuts methylated GATCs, coupled with PCR analysis or microarrays allows the generation of genome-wide DNA-binding maps of the protein.

Figure 1.7. Illustration of DamID principle.



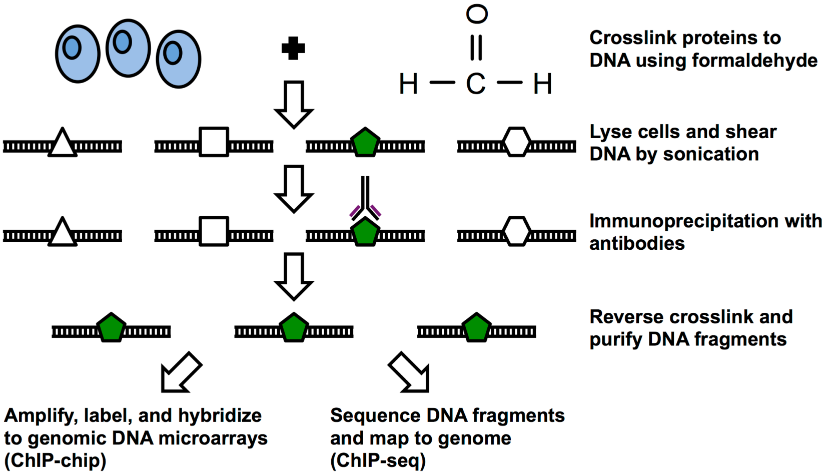
1.3.3 High-throughput methods

The previous methods are generally low throughput and laborious and what is more important, they generate consensus DNA-binding sites of low resolution due to the limited number of DNA sequences that can be tested in each experiment. In contrast, in the last decade, several methods have appeared able to characterize PDIs in a high-throughput manner.

ChIP-chip/ChIP-seq

Chromatin immunoprecipitation (ChIP) has been extensively used to characterize PDIs in vivo. This technique uses chemicals such as formaldehyde to crosslink proteins and DNA that are in direct contact, which is followed by cell lysate and DNA fragmentation by sonication. Then, the whole cellular content is immunoprecipitated in order to capture the protein of interest with specific antibodies, together with any crosslinked DNA fragment. After reversing the

Figure 1.8. Overview of ChIP-chip and ChIP-seq workflows.

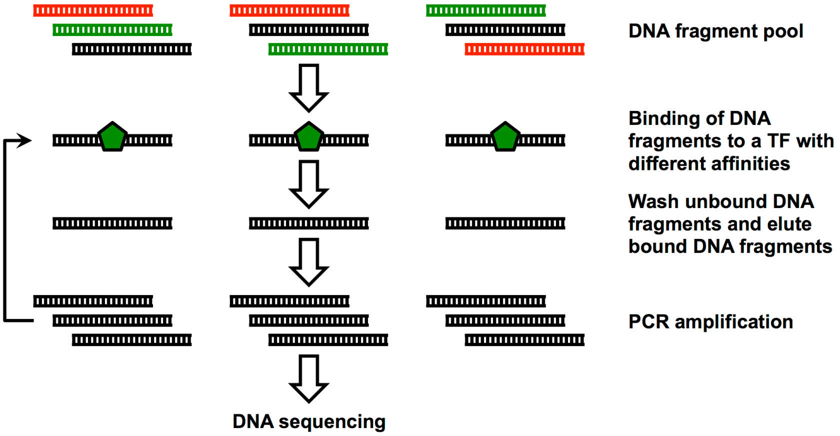

O
H — C — H
Crosslink proteins to DNA using formaldehyde
Lyse cells and shear DNA by sonication
Immunoprecipitation with antibodies
Reverse crosslink and purify DNA fragments
Amplify, label, and hybridize to genomic DNA microarrays (ChIP-chip)
Sequence DNA fragments and map to genome (ChIP-seq)

crosslink, the bound DNA is amplified and characterized in various ways such as microarrays (ChIP-chip) (160) or high-throughput sequencing (ChIP-seq) (161).

SELEX

The systematic evolution of ligands by exponential enrichment (SELEX) (162) is the main experimental approach used in the construction of the TRANSFAC database (163). It involves the incubation of a protein of interest with a library of random DNA

Figure 1.9. Overview of SELEX workflow.

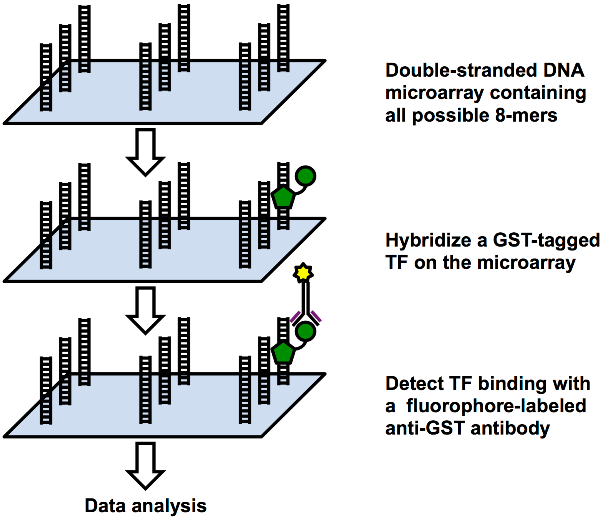

DNA fragment pool
Binding of DNA fragments to a TF with different affinities
Wash unbound DNA fragments and elute bound DNA fragments
PCR amplification
DNA sequencing

sequences. Then, bound and unbound DNA sequences are separated and bound DNA is further amplified by PCR and incubated with the protein. The whole procedure is repeated several times, which makes its application difficult. Moreover, it is also difficult to obtain TF motifs with high resolution, as SELEX typically yields between 20 and 70 binding sites. The introduction of serial analysis of gene expression (SAGE) overcomes these limitations (164). In SELEX-SAGE, radiolabeled oligonucleotides are used to monitor the binding conditions thus preventing the selection of only high-affinity binding sites. Moreover, the inclusion of SAGE increases the sequencing throughput by concatenation of the DNA sequences obtained during SELEX rounds.

Protein binding microarray

In a protein binding microarray (PBM) experiment, a recombinant protein of interest (GST-tagged) is probed against an array of double-stranded DNA and labeled with a fluorescent anti-GST antibody (165). The resulting intensity signals are used to determine the DNA-binding specificities of the given protein (the higher the intensity the stronger the protein is bound to DNA). One of the main advantages of the method is that it exhaustively explores the

Figure 1.10. Overview of PBM workflow.

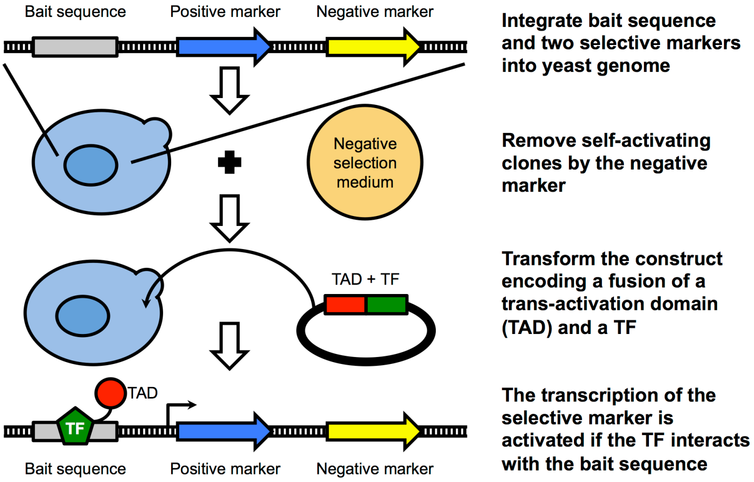

Double-stranded DNA microarray containing all possible 8-mers
Hybridize a GST-tagged TF on the microarray
Detect TF binding with a fluorophore-labeled anti-GST antibody
Data analysis

DNA space thus providing highly accurate consensus binding sites. The microarray is designed to contain all non-palindromic DNA sequences of 8 base pairs (8-mers) in at least 32 different spots (16 for palindromic 8-mers). This redundancy allows a robust estimation of the relative preference of the protein for every 8-mer. The binding site of the protein is reported as a PWM, which is derived from the top scoring 8-mers (166).

Yeast one-hybrid

Both SELEX and PBM approaches are useful for identifying the preferred binding sites of a given protein but, if the aim of the experiment is the opposite, to identify proteins that can specifically interact with a given DNA sequence, the yeast one-hybrid (Y1H) (167) or the protein microarray (171) methods are more indicated. Y1H exploits the fact that DNA-binding proteins, such as yeast Gal4, have a modular structure comprising a DNA-binding domain and a trans-activation domain (see "Structure of transcription factors" in section 1.2.2). In this method, a DNA fragment of interest (bait) is cloned upstream of two different reporters, *HIS3* and *LacZ*, which are then integrated into the genome of a yeast strain. Moreover, plasmids expressing a TF (prey) fused with the trans-activation domain of Gal4 are introduced into the yeast strain.

Figure 1.11. Overview of Y1H workflow.


Bait sequence
Positive marker
Negative marker
Integrate bait sequence and two selective markers into yeast genome
Negative selection medium
Remove self-activating clones by the negative marker
TAD + TF
Transform the construct encoding a fusion of a trans-activation domain (TAD) and a TF
TAD
TF
The transcription of the selective marker is activated if the TF interacts with the bait sequence
Bait sequence
Positive marker
Negative marker

If the prey interacts with the bait, the trans-activation domain of Gal4 activates the reporter expression regardless of whether the TF is an activator or repressor *in vivo*. The *HIS3* reporter overcomes 3AT inhibition allowing the yeast strain to grow, while the *LacZ* reporter produces a blue compound from X-gal. The sequencing of the plasmid in each of the blue colonies reveals the identity of the TFs that can bind the DNA fragment of interest.

Bacterial one-hybrid

In a bacterial one-hybrid (B1H) experiment, the TF of interest is expressed as a fusion to a subunit of RNA polymerase. In parallel, a library of randomized oligonucleotides with potential TF-binding sites is cloned into a separate vector that contains the selectable genes *HIS3* and *URA3*. If the TF (bait) binds a potential binding site (prey) *in vivo*, it will recruit RNA polymerase to the promoter and activate the transcription of the reporter genes in that clone. The two reporter genes, *HIS3* and *URA3*, allow for positive and negative selections, respectively. At the end of the process, positive clones are sequenced and examined with motif-finding tools in order to produce a PWM of the binding site (168).

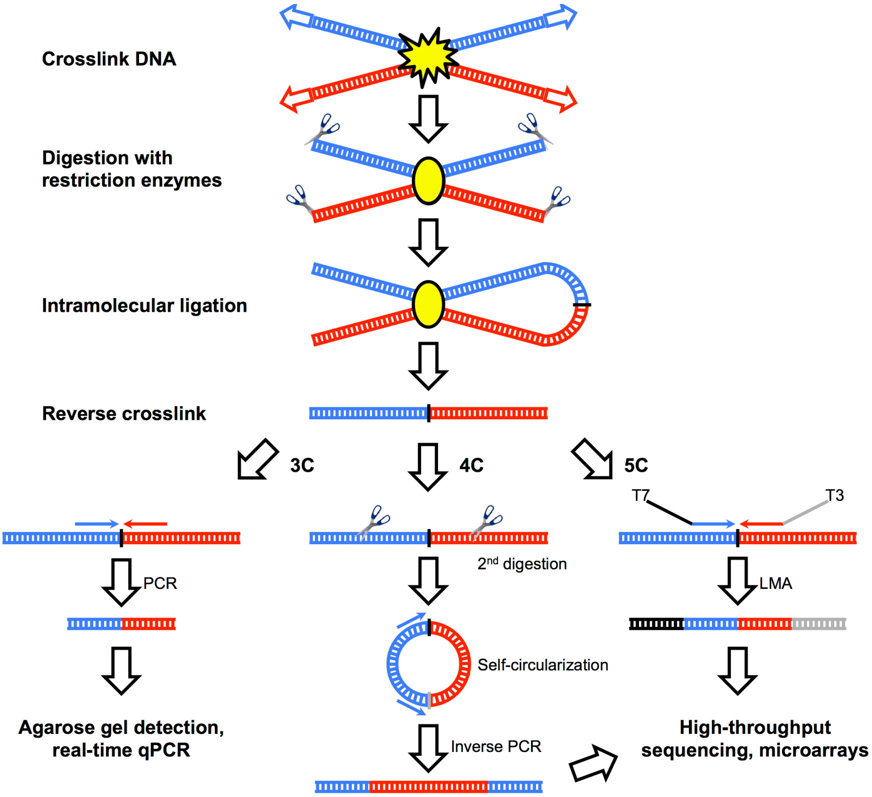
Protein microarray

Unlike PBMs, in a protein microarray experiment, a fluorescent-labeled DNA motif is probed against thousands of individually purified proteins on a glass. This approach, as in the case of Y1H, allows the identification of TFs that bind the promoter region of a gene of interest, and has been used to profile the protein-DNA interactome in human (169).

Chromosome conformation capture

The previous methods are indicated for the characterization of PDIs. However, studying the structural properties and spatial organization of chromosomes is also important to understand the regulation of gene expression. For example, some chromosomal regions can fold in order to bring an enhancer and its associated TFs within close proximity of a gene promoter (170). Chromosome conformation capture (3C) allows the analysis of such interactions *in vivo* (171). The whole genome is crosslinked with formaldehyde and digested with a restriction enzyme. This is followed by ligation at very low DNA concentrations to favor intramolecular ligation of crosslinked fragments over random fragments, whose ligation is intermolecular. The crosslink is then reversed and individual ligation products are finally detected and quantified by agarose gel detection or real-time qPCR using locus-specific primers. A disadvantage of 3C is the frequent random collisions of chromosomal regions to one another, which means that observed interactions between two regions are not always specific. Therefore, a specific interaction between two regions is only confirmed when the interaction occurs at a higher frequency than with neighboring DNA.

Figure 1.12. Overview of 3C, 4C, and 5C workflows.



An evolution of the method, circularized 3C (4C), has a significant advantage because only one of the two crosslinked fragments needs to be known (172,173). After the reversal of the crosslink, DNA fragments are digested a second time with a different restriction enzyme, which results in smaller fragments with restriction ends that differ from the previous restriction sites. This second digestion favors self-circularization of DNA that is further ligated. Circular DNAs are then amplified with inverse PCR, using primers designed against the outer restriction sites, and quantified by high-throughput sequencing or microarrays.

Another 3C-based method, the carbon-copy or 5C, allows the parallel analysis of interactions between many chromosomal regions (174). In 5C, after the crosslink is reversed, ligation-mediated amplification (LMA) is performed using multiplex universal primers as T7 and T3 fused to the ligation junction sequences (*i.e.* half the site recognized by the restriction enzyme). The primers anneal to the digested fragments and are ligated with a DNA ligase. The ligated primers are finally used as templates to further amplify and analyze the fragments *via* high-throughput sequencing or microarrays.

MNase-, DNase-, and other restriction enzyme-based methods

Also, the knowledge of the precise nucleosome locations in the genome is key to understand how gene regulation. Digestion of chromatin by micrococcal nuclease (MNase) provides information about nucleosome positioning along DNA strands (175). In this method, permeabilized cells are exposed to MNase in the presence of a divalent cation, which makes double-stranded cuts between nucleosomes. Treating the chromatin with very high concentrations of MNase yields mononucleosome-length DNA predominantly, while using lower concentrations of the enzyme generates one double-stranded cut at intervals of 10 to 50 nucleosomes, depending on the concentrations of both the enzyme and chromatin. MNase can also make single-stranded DNA cuts at the sites of histone octamers, and thus, the mapping of nucleosome occupancy is usually performed with native double-stranded DNA.

Although DNase I has traditionally been used to predict PDIs *in vitro*, its ability to digest accessible DNA in internucleosomal regions has been exploited to probe nucleosome occupancy *in vivo* (176). The digestion of nuclear chromatin by DNase I produces DNA fragments smaller than 500 base pairs that, in theory, allows to distinguish DNA fragments bound by either TFs or nucleosomes (177). The use of DNase over MNase has the advantage of coupling both TF occupancy and nucleosome positioning (178). However, a recent article shows that MNase coupled to immunoprecipitation of native chromatin (N-ChIP) can be used to generate high-resolution maps of TF binding sites on native chromatin (179).

A recently published method, FIREWACh (180), uses restriction enzymes in permeabilized cells to isolate nucleosome-free regions of DNA. The resulting fragments are amplified using ligation-mediated PCR with a complementary flanking adaptor primer, and inserted within the lentiviral reporter plasmid of FpG5, upstream of GFP-coding sequences. Plasmids are then transfected to another sample of the original cells and GFP$^+$ cells are selected using fluorescence-activated cell sorting cytometry. Further amplification with PCR and sequencing of the plasmids in GFP$^+$ cells results in the identification of transcriptional regulatory modules.

1.3.4 High-resolution methods

A deepen comprehension on how proteins and DNA interact can be achieved using methods that characterize protein-DNA complexes at atomic resolution. In the past decade, advances in the structural characterization of PDIs have accelerated the field thus facilitating the determination of large protein-DNA complex structures such as the formation of chromatin fibers due to tetranucleosomes (38), or even the encircling of DNA occurring during gene transcription and

DNA replication by sliding clamps (181,182). Still, the structural determination of protein-DNA complexes remains challenging, as suggested by the limited number of protein-DNA complexes stored in the Protein Data Bank (PDB) (183) (see Figure 1.13). For example, less than 1% of structures of human proteins in the PDB correspond to PDIs[5]. There are a few experimental techniques including modeling procedures that can be used in order to elucidate the structure of PDIs at atomic resolution.

X-ray crystallography

X-ray crystallography is the most widely used experimental method for determining the structure of large biomolecules. In fact, more than 90% of released PDIs in the PDB[5] have been solved using this technique. It allows the three-dimensional (3D) description at atomic resolution of a crystallized macromolecule, in which their atoms cause a beam of incident X-rays to diffract into many specific directions. The produced diffraction map contains information about the electron density of the macromolecule, which allows the elucidation of their atom positions and chemical bonds. However, not all the contacts observed in a crystal are biologically relevant (184), and the crystallization conditions may not represent exactly those of an *in vivo* environment.

Nuclear magnetic resonance spectroscopy

In nuclear magnetic resonance (NMR), the macromolecule of study is kept in solution, which is a more natural environment. The sample is placed under a strong magnetic field and short frequency pulses are used to excite the nucleus of the macromolecule. This allows the detection of different chemical shifts for each nucleus of the macromolecule, which depend on their chemical environment. The different radio frequency pulses and the analysis of the chemical shifts that they produce in the different nucleus are used to determinate the distances between the different atoms and generate a 3D-structural model at atomic resolution of the macromolecule.

[5] Last accessed: July 2014.

Moreover, amide hydrogen exchange experiments using deuterated water can be used to identify changes in solvent accessibility of interfacial residues and to provide additional evidence of the protein-DNA interface (185). However, the application of NMR is limited by the size of the complex of interest (186).

Small-angle scattering and cryo-EM tomography

Small-angle scattering, either using X-rays or neutron beams, is a promising alternative technique for the structural characterization of protein-DNA complexes (187). The macromolecule is exposed to X-rays or neutron beams and a detector registers the scattered radiation. Then, the X-ray or neutron scattering curve (intensity versus scattering angle) is used to create a low-resolution model of the system (around 15 Å). In contrast to the previous structural methods, small-angle scattering experiments can be done in a few days. In addition, the fact that a crystalline sample is not needed allows the study of the dynamic properties of the macromolecule in solution, which is a more realistic environment. In cryo-electron microscopy (cryo-EM) tomography, the macromolecule is observed at cryogenic temperatures by an electron microscope, which uses a beam of electrons to create an image. However, highly dynamic systems as protein-DNA complexes (188) difficult the interpretation of density maps, thus affecting the structural resolution of this technique (around 15 Å). Still, both small-angle scattering and cryo-EM data provide information about the shape and size of the macromolecule, which can be exploited in computational modeling to solve structures at atomic resolution of large macrocomplexes (revised in section 3.2.1).

Computational modeling

Although modeling procedures involve the use of computers, I will revise them in this section due to their ability to yield 3D models of PDIs at atomic resolution. The most common way of modeling PDIs is *via* comparative modeling (revised in (190)). This approach can only be applied as long as there is a homologous structure of the interaction. Then, applications such as MODELLER (191) can be used to model the protein of interest. Moreover, the modeling of DNA requires the use of specialized software, as for example 3DNA (192). Occasionally, PDI models can be constructed by superimposition of the structure/model of the unbound protein over the structure of a homologous protein-DNA complex (193) using structural alignment programs like TM-align (194).

In general, homology modeling is limited by those homologs whose structure is too remote to help assigning the correct fold. However, two proteins can have similar structures even if they share little sequence similarity (*i.e.* the twilight zone) (195). Alternatively, docking procedures can be used if there are no available structures of the interaction, but the structure of both the protein and DNA are known or can be modeled (196–200). This subject is revised in detail in section 3.2.1 of this thesis.

1.3.5 Computational methods

Given the complexity of genome regulation observed in eukaryotic organisms, it is likely that the relevant details about most DNA-binding proteins and/or their binding sites still remain unknown for most of them. Despite the recent advances described in this section, experimental characterization of PDIs is a laborious and difficult process and as a result, only a small fraction of eukaryotic TFs has been profiled (201). Therefore, it is necessary the development of computational approaches for rapid and accurate mapping of PDIs, especially to complement experimental methods. Prior to this PhD thesis, there existed several computational methods that address three closely related problems regarding protein-DNA association: predict DNA-binding proteins; infer protein-DNA interfaces; and characterize protein DNA-binding sites. A summary of these tools is provided in Tables 1.3, 1.4 and 1.5. A less explored topic is the prediction of TFs that specifically recognize and bind a given DNA sequence. This has only been achieved with a certain degree of success in the case of zinc finger endonucleases (202).

Identification of DNA-binding proteins

The different approaches to identify DNA-binding proteins can be divided into two groups: sequence-based and structure-based. On the one hand, sequence-based methods take into account different features of the protein such as its amino acid composition (203,204) or evolutionary profiles (205,206). However, these methods cannot discriminate DNA-binding from RNA-binding proteins (207). On the other hand, structure-based methods (208–210) normally yield better predictions than sequence-based ones.

Prediction of protein-DNA interfaces

In a similar way, the observed amino acid conservation in protein-DNA interfaces (211) has been exploited to train machine learning approaches in order to predict the DNA-binding residues of a protein (212–216). However, these methods have two main drawbacks: they assume the interaction of the protein with DNA and they tend to overpredict positively charged residues. Again, if the structure of the protein is known, it can be used to calculate the accessible surface area of each amino acid, or to extract structural properties, which results in better predictions of the amino acids that contact DNA (217–219).

Characterization of protein-DNA binding sites

A well-established procedure to characterize protein DNA-binding sites is to search with a motif discovery algorithm for over-represented DNA sequences in the promoter regions of genes known to be regulated by that protein (220). However, the success of such approaches requires the availability of enough sequences for pattern discovery, which often is not the case. Moreover, for homeodomain and zinc finger proteins, which happen to be the largest families of TFs, there are models that allow the prediction of

Table 1.3. Tools for discriminating DNA-binding proteins.

Name	Features	Methods used	Query
DBD-Threader (209)	All-atom, distance-dependant, biochemical features	Statistical potentials, structural superimposition, threading	Protein sequence
DNABinder (205)	Amino acid properties, evolutionary profiles	Support vector machine	Protein sequence
iDNA-Prot (221)	Amino acid composition from grey model	Random forest	Protein sequence
ProteDNA (222)	Secondary structure	Support vector machine	Protein sequence
iDBPs (206)	Conservation, evolutionary information, dipole moment, secondary structure, electrostatics, amino acid sequence	Random forest	Protein structure
DBD-Hunter (208)	All-atom, distance-dependant, biochemical features	Statistical potentials	Protein structure
SPOT (210)	All-atom, distance-scaled, finite, ideal-gas reference state	Statistical potentials	Protein structure

Table 1.4. Tools for characterizing protein-DNA binding sites.

Name	Features	Methods used	Query
3DTF (223)	All-atom, distance-dependant	Statistical potentials	Protein-DNA complex structure
DBD2BS (224)	All-atom, distance-scaled, finite, ideal-gas reference state	Statistical potentials	Protein-DNA complex structure
DDNA (225)	All-atom, distance-scaled, finite, ideal-gas reference state	Statistical potentials	Protein-DNA complex structure
PiDNA (226)	All-atom, distance-dependant	Statistical potentials	Protein-DNA complex structure

Table 1.5. Tools for predicting protein-DNA interfaces.

Name	**Features**	**Methods used**	***Query***
BindN (212)	Side chain pK_a, hydrophobicity, molecular mass	Support vector machine	Protein sequence
BindN+ (215)	Biochemical features, evolutionary information	Support vector machine	Protein sequence
BindN-RF (227)	Side chain pK_a, hydrophobicity, molecular mass, conservation, biochemical features, PSSM	Random forest	Protein sequence
DBindR (228)	Evolutionary profiles, secondary structure, orthogonal binary vector information	Random forest	Protein sequence
DBS-Pred (203)	Sequence information, solvent accessibility, secondary structure	Neural network	Protein sequence
DBS-PSSM (229)	Evolutionary profiles	Neural network	Protein sequence
DP-Bind (213)	Amino acid properties, evolutionary profiles	Support vector machine	Protein sequence
NAPS (214)	Amino acid properties, evolutionary profiles	Decision tree algorithm	Protein sequence
metaDBSite (216)	Sequence information, results from 6 other servers	Support vector machine	Protein sequence
PreDNA (230)	Evolutionary profiles, structural geometry	Support vector machine	Protein sequence
ProteDNA (222)	Secondary structure	Support vector machine, secondary structure alignment	Protein sequence
DISPLAR (217)	Position-specific sequence profiles, solvent accessibility	Neural network	Protein structure
DNABINDPROT (218)	Fluctuations of residues in high-frequency modes	Gaussian network model	Protein structure
DR_bind (219)	Solvent accessibility, conservation, structural geometry, electrostatics		Protein structure

the PWM of a protein just from its sequence (231–233). Finally, if not only the structure of the protein is known but also the structure of its complex with DNA, then, statistical potentials can be applied in order to predict the DNA targets of a protein and generate a PWM (193,223,225,226,234–237).

Databases of protein-DNA interactions

Information stored in databases of PDIs and/or protein-DNA complex structures can be used as a complement to computational methods. For example, TRANSFAC (163) is a manually curated database of eukaryotic TFs, and their experimentally determined binding sites and binding profiles. A similar database, JASPAR (238), only contains open access data and is non-redundant.

Some databases are very specialized in their content. For example, the databases ChIPBase (239) and UniPROBE (240) only contain PDIs determined by either ChIP-chip or PBMs, respectively. There are other databases that are specie-specific, such as YEASTRACT (241) and YeTFaSCo (242) in yeast, REDFly (243) and OnTheFly (244) in fruit fly, or hPDI (245) in humans. Even there are databases exclusively dedicated to certain types of TFs, as it is the case of the homeodomain (246) and zinc finger families (247). Also, the thermodynamics of protein-DNA complexes, including the affinity of binding upon complex formation, the strength of the interaction, or the effects of mutations in amino acids or nucleotides on the binding specificity, can be found in ProNIT database (248).

The available protein-DNA structural data stored in the PDB have been collected in several databases such as NPIDB (249), 3D-footprint (250), BIPA (251), TFinDit (252), or footprintDB (253) (the last two databases exclusively contain structures of bound and unbound TFs). Some databases go one step further and classify different types of contacts (254,255) or features (256) that can be extracted from the previous structures.

Statistical potentials derived from 3D-structures

Most of the previous methods that exploit 3D-structures of PDIs are based on statistical potentials. In general, statistical potentials are energy functions derived from the analysis of known structures in the PDB. These functions are used to score all contacts observed in a protein structure or complex (a PPI or a PDI) and provide an approximation of its free energy (the lower, the better). The section 3.2.1 of this thesis is exclusively dedicated to statistical potentials: what they are, how they are derived for PDIs, and an extensive case study.

1.4 Motivation of this thesis

Experimental data on TF-binding sites and nucleosome occupancy shed light on our understanding on how genes are regulated as such data continue to accumulate. However, we are still far from having a full picture on how the different DNA regions in a cell interact with each other, directly or indirectly through their RNAs and protein expression products. In part, it is because the experimental characterization of PDIs is a laborious and difficult process. In this context, computational techniques, either sequence- or structure-based, play an indispensable role to help disclose gene regulation.

On the one hand, sequence-based approaches have focused on characterizing DNA-binding proteins and their interfaces. However, little effort has been made towards predicting TF-binding sites just from their sequence. A classic approach is the use of homology-based tools for transferring annotation between similar TFs (257). Still, TF-binding preferences, even between closely related TFs, are usually determined by a few key amino acids (258), which blurs the predictions of such methods. As I have already shown in section 1.2.3, TFs act cooperatively, they need to interact with other proteins in order to effect their function (133). Moreover, PPI data has been previously applied with success to predict distant related proteins (259) or to infer enzyme function (260) (I review this subject in (151); see Appendix 1). Thus, PPIs can be exploited to improve homology-based methods. In section 3.1 of this thesis, I show the prediction of redundant TFs (*i.e.* TFs that can bind to the same binding sites) by combining homology-based tools and PPI data. The approach was further automated and generalized to infer the fold of remote homologs (195).

On the other hand, structure-based approaches have also been used to predict TF-binding sites. These methods are based on statistical potentials and, although they are a good alternative to experimental techniques, their application is not exempt from limitations. One of them is the scarcity of protein-DNA complex structures available in the PDB. To avoid any bias, statistical potentials are usually derived from a non-redundant data set of structures, which can generate statistical potentials suffering from low-count and at the same time low diversity of binding patterns. Another setback is that statistical potentials assume the contribution of the different DNA base pairs to the binding energy of the complex is independent from each other, which is incorrect (261). Therefore, there is still room for improvement in the area. In sections 3.2.1 and 3.3.1 I tackle both problems by 1) describing statistical potentials for contacts between amino acids and dinucleotides (*i.e.* pairs of consecutive nucleotides along the DNA sequence) and 2) incorporating experimental data from PBMs to statistical potentials.

Finally, as stated in section 1.3.5, the prediction of the best TF for a given DNA sequence has only been achieved in the case of zinc fingers endonucleases (202). In section 3.3.1 I show the application of statistical potentials regarding this subject.

2. Objectives

“Learn from yesterday, live for today, hope for tomorrow. The important thing is to not stop questioning.”
-Albert Einstein

This thesis aims to fulfill the following objectives:

1. To study redundancy in gene regulatory networks caused by transcription factors.
2. To predict DNA sequences that can be bound to a given transcription factor.
3. To infer transcription factors that can bind to a given DNA sequence.

The achievement of these goals comprises several milestones:

- To exploit homology-based tools together with protein-protein interactions for remote homology prediction of function and fold (see section 3.1.1 and 3.1.2).
- To revise current state-of-the-art statistical methods to infer protein-DNA interactions (see section 3.2.1).
- To describe statistical potentials for protein-DNA interactions (see section 3.2.1 and 3.3.1).
- To incorporate protein binding microarray data to statistical potentials (see section 3.3.1).
- To develop an automated modeling pipeline for transcription factor-DNA complexes (see section 3.3.1).

3. Results

“If your experiment needs a statistician, you need a better experiment.”
-Ernest Rutherford

3.1 Studying redundancy in transcription factors

The complementarity between gene expression and protein-DNA interaction data has led to several successful models of biological systems. However, several studies in multiple species have raised doubts about the relationship between these two datasets. These studies have shown that the overwhelming majority of genes bound by a particular TF are not affected when that factor is knocked out. I hypothesize knockouts could be compensated by redundant TFs (*i.e.* a homolog that could replace the deleted TF). To further prove this hypothesis, I combine homology-based tools together with PPI data to predict remote homology in TFs. I also automate the approach to detect remote homology and generalize it to predict protein fold in ModLink+.

Manuscripts presented in this section:

Gitter, A., Siegfried, Z., Klutstein, M., **Fornes, O.**, Oliva, B., Simon, I., & Bar-Joseph, Z. (2009). **Backup in gene regulatory networks explains differences between binding and knockout results.** *Molecular Systems Biology*, 5, 276.

Fornes, O., Aragues, R., Espadaler, J., Marti-Renom, M.A., Sali, A., & Oliva, B. (2009). **ModLink+: improving fold recognition by using protein-protein interactions.** *Bioinformatics,* 25(12), 1506-1512.

3.1.1 Backup in gene regulatory networks explains differences between binding and knockout results

My contribution to this manuscript was to detect homologous TFs and calculate the percentage of shared PPI for each of them.

Gitter A, Siegfried Z, Klutstein M, **Fornes O**, Oliva B, Simon I et al. Backup in gene regulatory networks explains differences between binding and knockout results. Mol Syst Biol. 2009; 5: 276. DOI: 10.1038/msb.2009.33

3.1.2 ModLink+: improving fold recognition by using protein-protein interactions

Fornes O, Aragues R, Espadaler J, Marti-Renom MA, Sali A, Oliva B. ModLink+: improving fold recognition by using protein-protein interactions. Bioinformatics. 2009 Jun 15;25(12):1506-12. doi: 10.1093/bioinformatics/btp238

3.2 Statistical potentials for protein-DNA interactions

In order to perform their function, proteins need to interact with each other as well as with other biomolecules such as DNA or RNA. Therefore, to fathom the function of a protein, we need to know with whom it can interact as well as the atomic details of such interactions (*i.e.* the structure of the complex). However, the total amount of protein interactions with an experimentally determined 3D-structure is small. Therefore, computational modeling is key to fill this gap. Protein interactions can be modeled using as templates the interactions of homologous proteins, if the structure of the complex is known, or using docking procedures. No matter the approach used, the estimation of the quality of the produced models is essential. In this section, I revise several applications of statistical potentials: from the assessment of models to the ranking of docking possess, passing through the prediction of interfaces and PDIs. Moreover, I further show how to derive split-statistical potentials for PDIs together with a case study.

Manuscript presented in this section:

Fornes, O., Garcia-Garcia, J., Bonet, J., & Oliva, B. (2014). **On the Use of Knowledge-Based Potentials for the Evaluation of Models of Protein-Protein, Protein-DNA, and Protein-RNA Interactions.** *Advances in Protein Chemistry and Structural Biology,* 94, 77-120.

3.2.1 On the use of knowledge-based potentials for the evaluation of models of protein-protein, protein-DNA and protein-RNA interactions

Fornes O, Garcia-Garcia J, Bonet J, Oliva B. On the use of knowledge-based potentials for the evaluation of models of protein-protein, protein-DNA, and protein-RNA interactions. Adv Protein Chem Struct Biol. 2014;94:77-120. doi: 10.1016/B978-0-12-800168-4.00004-4

3.3 Prediction of protein-DNA interactions

Statistical-based approaches have been exploited in different works to predict TF-binding sites (see sections 1.3.5 and 3.2.1). However, the application of these methods is limited and can be improved. In this context I present SHAITAN, a homology-based approach that combines structural information and protein binding microarray data for the annotation of TF-binding sites and the discovery of TFs able to bind to specific DNA regions.

Manuscript presented in this section:

Fornes, O., Weirauch, M.T., Hugues, T.R., & Oliva, B. **SHAITAN: On the prediction of protein-DNA interactions with a homology-based approach** (*To be submitted*).

3.3.1 SHAITAN: On the prediction of protein-DNA interactions with a homology-based approach

<u>Fornes, O.</u>, Weirauch, M.T., Hugues, T.R., & Oliva, B. **SHAITAN: On the prediction of protein-DNA interactions with a homology-based approach** (*To be submitted*).

SHAITAN: On the prediction of protein-DNA interactions with a homology-based approach

Oriol Fornes[1], Matthew T. Weirauch[2,3], Timothy R. Hughes[3,4] and Baldo Oliva[1,*]

[1]Structural Bioinformatics Lab. (GRIB), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Catalunya, Spain
[2]Center for Autoimmune Genomics and Etiology (CAGE) and Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA
[3]Banting and Best Department of Medical Research and Donnelly Centre, University of Toronto, Toronto ON M5S 3E1, Canada
[4]Department of Molecular Genetics, University of Toronto, Toronto ON M5S 1A8, Canada

[*]To whom correspondence should be addressed: baldo.oliva@upf.edu

ABSTRACT

The knowledge on transcription factor (TF) binding sites is key to understand gene regulation. However, the binding preferences for most of eukaryotic TFs are unknown. In this context, the development of computational tools as a complement to experimental procedures for characterizing TF-binding sites is foremost. In this work, we present SHAITAN, a homology-based approach that combines structural information and protein binding microarray data for the annotation of TF-binding sites and the discovery of TFs able to bind to specific DNA regions. SHAITAN was superior than three other state-of-art methodologies when benchmarked to predict the binding sites for 70 TFs from the DREAM5 Challenge. SHAITAN was also successful at identifying different combinations of specificity residues for determined DNA sequences in homeodomains. Thanks to an automated homology modeling pipeline for TF-DNA complexes, SHAITAN could be applied at large-scale in order to fill the existing gaps in gene regulatory networks.

INTRODUCTION

DNA-binding proteins play essential roles in the daily activities of cells. Among them, transcription factors (TFs) are the most widely studied. TFs are able to activate/repress gene transcription by binding to specific sites in enhancer or promoter regions adjacent to their regulated genes. The variability observed among DNA sequences recognized by a TF has driven to use probabilistic models, called position weight matrices (PWMs), in order to represent TF-binding sites. Specifically, researchers have exploited PWMs to search for novel targets of TFs (reviewed in (1)).

Characterizing TF-binding sites is foremost in order to understand how genes are regulated. In the past decade, the appearance of high-throughput techniques, including ChIP-chip (2) and ChIP-seq (3), protein binding microarrays (PBMs) (4), or the bacteria and yeast one-hybrids (5,6), have allowed the characterization of TF-binding sites at large-scale. However, experimental protocols are both laborious and difficult to apply, as it is suggested by the very small fraction of eukaryotic TFs that have been profiled (7). As an alternative, computational tools can be employed. A well-established procedure to infer the binding sites of a TF consists in searching the promoter regions of its regulated genes for over-represented sequence patterns using a motif discovery algorithm (8). Nevertheless, the success of this strategy requires the availability of enough sequences for pattern discovery, which is only possible for a small number of TFs. Another successful approach to predict a TF-PWM is the structural analysis of its complex with DNA using statistical potentials (revised in (9)). Briefly, given a TF-DNA complex structure, different DNA sequences are threaded and the binding energies of the resulting interactions are calculated. Best scoring sequences are then considered to be bound by the TF and incorporated into a PWM.

Although statistical potentials are a good alternative to infer TF-binding sites, their application still has some limitations. One of them is the lack of templates due to the small number of TF-DNA complex structures available in the Protein Data Bank (PDB) (10). To avoid any bias, statistical potentials are usually derived from a non-redundant (nr) dataset of PDB structures. This redundancy is generally removed on the TF side of the complex. Yet, TFs can recognize different binding sites and in addition, members of the same family of TFs can bind to distinct DNA sequences (11). For this reason, the removal of redundancy can generate statistical potentials suffering from low-count and at the same time low diversity of binding patterns. Another problem arises because statistical potentials assume the contribution of the different DNA base pairs to the binding energy of the complex is independent from each other, which is not true (12). AlQuraishi and McAdams addressed the coverage problem for the homeodomain family by combining TF-DNA structures with experimentally determined PWMs (13). The inclusion of PWM data adapted the statistical potential to the varying binding preferences of homeodomains for different sites. Still, they highlighted the use of PWMs cannot account for inter-position dependencies observed among base pairs.

In a recent work, we used statistical potentials of contacts between amino acids and dinucleotides (*i.e.* pairs of consecutive nucleotides along the DNA sequence) to address the recognition of DNA binding sites by TFs (9). We relied on a small set of nr TF-DNA complex structures to derive the statistical potentials. However, these potentials could not cover the whole contact spectrum and, as a result, the coverage of protein-DNA contacts was undermined. In this work, we present SHAITAN, Statistical-potentials using a Homology-based Approach to predict the Interactions between Transcription factors And Nucleotide sequences. SHAITAN is unique in that it integrates structural information and PBM data into three statistical potentials, thus bypassing the coverage problem. We further prove SHAITAN's ability to infer TF-binding sites as well as to detect TFs able to bind a particular DNA sequence. SHAITAN was superior than other state-of-art tools when benchmarked against 70 targets from the DREAM5 TF-DNA Motif Recognition Challenge (14). Besides, SHAITAN was successfully applied to predict the specific residues involved in the recognition of the DNA binding site for the homeodomain family of TFs.

RESULTS

When facing the problem of protein-DNA interactions we could address the following two questions: 1) what DNA binding sites were recognized by a TF; and 2) what TFs were able to bind a specific DNA sequence. In order to answer these questions, we could either choose the TF family or search over all possible TF conformations (*i.e.* all families). We tackled both challenges by using two different types of statistical potentials (see Methods): 1) specific for a particular TF family; and 2) general potentials that could be applied to all TF families. Thus, we considered deriving such potentials from two different sets of structures. These sets were constructed by removing similarity with two different thresholds: a strict cut-off, applicable to all structures, and a less restrictive cut-off that allowed the inclusion of TF-structures of the same family. The selected cut-offs were 35% and 70%, respectively, and the resulting non-redundant (nr) sets were named nr35 and nr70. These sets were further used to train SHAITAN statistical potentials E_{S3DC}, E_{S3DCdd} and E_{S3DCdi} under different conditions (see Methods).

Evaluation of SHAITAN statistical potentials

We tested different statistical potentials (*i.e.* E_{S3DC}, E_{S3DCdd} and E_{S3DCdi}) to discern positive from negative 8-mers (*i.e.* P8Ms and N8Ms, respectively) in the PBM data using a five-fold cross-validation approach (see Methods). For each evaluated potential, we calculated the area under the receiver operating characteristic curve (AUROC) as a measure of performance. However, as we were dealing with unbalanced data, the performance of AUPR (*i.e.* area under the precision recall curve) was more informative than other statistical measures (15).

Table 1 shows the resulting averaged AUROCs and AUPRs of SHAITAN using statistical potentials E_{S3DC}, E_{S3DCdd} and E_{S3DCdi} derived under different conditions. The use of PBM information to generate the statistical potentials improved the performance of SHAITAN with respect to potentials generated with only known structures. When using the non-specific general potentials with information from PBM, the AUPR was 1.5 times better than with information derived only from known structures. Besides, SHAITAN achieved

AUROCs over 0.9 when using family-specific potentials. In particular, for E_{S3DCdd}, SHAITAN achieved an AUPR more than 60 times better than expected by random.

Table 2 shows the averaged AUROCs and AUPRs of SHAITAN using family-specific potentials. We have to note that, with independence of the statistical potential used, the performance of SHAITAN on the C2H2 zinc finger family was not successful. Members of this family of proteins typically have multiple DNA-binding domains arranged in tandem arrays, with each domain binding to 3 or more bases, and with an offset between fingers of three bases (16). Therefore, when mapping PBM data from multi-fingered proteins with more than 3 fingers over PDB structures, we could not tell which of the fingers interacted with the probes in the PBM experiment. Furthermore, depending on the PDB structure selected as template (see Methods), the maximum number of consecutive zinc fingers we could model was either 5 (17) or 6 (18). Thus, when modeling C2H2 zinc finger proteins with more than 6 fingers, more than one model was possible and we could not ascertain which was the correct model of the interaction as occurring in the PBM. These drawbacks resulted in a source of noise when deriving family-specific potentials for this family (see Figure 1). Still, SHAITAN could achieve AUROCs larger than 0.9 for the majority of TF-families (20 out of 24). Furthermore, when using the statistical potential E_{S3DCdd}, AUPRs were more than 50 times better than expected by random for 70% of the families.

Finally, we computed the positive predictive value (PPV) of SHAITAN using family-specific potentials in order to select the best cut-off for each of them to distinguish P8Ms from N8Ms (see Figure 2). In all cases, SHAITAN achieved maximum PPVs at energies lower than -0.45 (higher than 0.45 when using E_{S3DCdi}). E_{S3DCdd} was the most successful statistical potential implemented in SHAITAN, as shown by the statistical measures obtained in the different tests. Surprisingly, E_{S3DCdi}, a distance independent potential, was better than the original E_{S3DC} from which E_{S3DCdi} was derived, with PPVs around 20%. So, the best statistical potentials in SHAITAN were, by decreasing order, E_{S3DCdd}, E_{S3DCdi} and E_{S3DC}.

Prediction of DREAM5 PWMs

We tested the capacity of SHAITAN family-specific potentials E_{S3DC}, E_{S3DCdd} and E_{S3DCdi} to report TF-PWMs (see Methods). We compared these potentials with three state-of-art servers available online: 3D-footprint (19), 3DTF (20) and PiDNA (21). We used as benchmark the PWMs of 83 mouse TFs from the DREAM5 TF-DNA Motif Recognition Challenge (14). The three web servers require a TF-DNA complex structure, in PDB, format as input, and return the predicted TF-PWM as output. Nevertheless, most DREAM5 targets were not present in our set of known structures (see Methods). In all targets tested, the TF-DNA complex was modelled as shown in Figure 3. In total we obtained a TF-DNA complex structure, either directly from the PDB or modelled by our approach, for 70 DREAM5 targets.

Figure 4 shows the quality of the predictions by means of comparing the PWMs created by 3D-footprint, 3DTF, PiDNA and SHAITAN to the DREAM5 PWMs. Comparisons were made using Tomtom (22) as distributed in the MEME package (23). Tomtom yields the similarity between a pair of PWMs in form of a p-value. A prediction was considered good (green), mediocre (yellow), or bad (red) if the Tomtom p-value was $\leq 10^{-3}$, between 1 and 10^{-3}, or 1, respectively. Moreover, we considered a DREAM5 target as either easy or difficult to be predicted if, according to FIMO (24), the PWM bore some resemblance with the nucleotide sequence of the target TF-DNA complex. Additionally, we tested SHAITAN statistical potentials E_{S3DC}, E_{S3DCdd} and E_{S3DCdi} derived without using PBM data specific of the DREAM5 targets (which was available in the real challenge).

PiDNA and SHAITAN with E_{S3DCdd} were applied to more than half the DREAM5 targets. They produced good results for 25 and 31 targets, respectively. Although the majority of good predictions were achieved for easy targets (20 out of 25 and 31, respectively), the capability of SHAITAN to yield good PWM predictions among the difficult targets was remarkably better than the rest of methods. When evaluating their performance on individual families, PiDNA and SHAITAN, using statistical potentials E_{S3DC} and E_{S3DCdd}, predicted good PWMs for TFs belonging to 7 different families. Surprisingly, SHAITAN with E_{S3DCdi} also achieved good predictions on members from 6 different families, while 3D-footprint and 3DTF were far behind reporting only good PWMs for 4 and 3 families, respectively. Still, good PWMs could not be obtained for the GATA, IRF, MBD and POU TF families with any of these approaches. As expected, the coverage on the C2H2 zinc finger family was very limited, especially in the case of multi-fingered proteins (*i.e.* from Zbtb1 to Zscan20).

Prediction of homeodomain protein-PWMs

Finally, we further investigated the ability of SHAITAN to predict putative TFs able to bind a DNA sequence of interest (see Methods). As a case study, we selected the homeodomain family. Homeodomains consist of three α-helices comprising approximately 60 amino acids. Upon binding, the third helix of the domain (*i.e.* the recognition helix) is accommodated in the major groove and forms contacts with DNA bases. The specificity of the interaction resides on amino acids 47, 50 and 54, which are located in the center of the recognition helix (25). Berger *et al.* revealed the binding preferences for the majority of mouse homeodomain proteins using PBMs, and found certain combinations of specificity amino acids that allowed homeodomains to bind predominantly to determined 8-mers (26).

In Figure 5, we show for different DNA sequences the results to predict the 3 specificity residues located in the recognition helix of the homeodomain family. We also show the PWM logos of the region comprising residues 47-54 predicted by SHAITAN (see Methods). As we can see, SHAITAN was able to correctly predict the homeodomain specificity amino acids for DNA sequences scored under the cut-off of -0.45 (*i.e.* marked with a green tick). In contrast, unreliable predictions failed to provide the correct specificity residues (*i.e.* energies over the -0.45 cut-off and marked with a red cross).

DISCUSSION

TF-binding specificities represent the cornerstone on functional genomics and the analysis of gene regulation. Still, only about 1% of eukaryotic TFs have been profiled using "omics" techniques (7). In this regard, the development of computational tools as a complement to experimental procedures is foremost. In the recent years, several approaches based on statistical potentials have emerged with the objective to annotate TF-PWMs (revised in (9)), but their application is complicated for the non-specialist, since they require the knowledge of the structure (or model) of the TF of interest in complex with DNA.

In this work, we have presented SHAITAN, Statistical-potentials using a Homology-based Approach to predict the Interactions between Transcription factors And Nucleotide sequences. SHAITAN integrates structural information and PBM data into three different statistical potentials. Moreover, we have demonstrated SHAITAN is superior than other current state-of-art methods available online on the prediction of TF-PWMs. Additionally, we have supplied a case study in which we show the application of SHAITAN to detect the best TF able to bind a particular DNA sequence.

However, it has to be noted the reliability of the predictions for certain TF families was dubious, especially for the C2H2 zinc finger family. As shown in Figure 1, PBM data is difficult to interpret for some members of this family, being impossible to model the correct conformation of the binding complex, which in turn affects negatively the performance of the method. However, we suggest this problem could be solved by using available information on individual zinc fingers (27) instead of PBM data.

Thanks to an automated pipeline for modeling TF-DNA complexes, SHAITAN was easily converted into a large-scale platform to annotate TF binding sites, thus providing the scientific community with a powerful tool that can be applied to fill the existing gaps in gene regulatory networks (28) or propose putative transcription factors that interact with a specific DNA binding site. Furthermore, SHAITAN can be applied to redesign the best protein sequence of a TF belonging to a particular family, or to identify the best specific residues able to bind a determined DNA sequence. An example was provided for the homeodomain family, for which we covered many of the known experimental TF targets. Last but not least, we defined the limits of a reliable prediction: as important as it is to predict the possibility a TF binds a particular DNA binding site, it is also relevant to know whether we can rely on the prediction or not.

METHODS

Experimental data and software

We retrieved from the PDB a total of 702 TF-DNA complex structures annotated in the TFinDiT (29) depository (April 2014 release). Moreover, we downloaded PBM data from various studies (14,26,30–42) available in the catalog of inferred sequence binding preferences (Cis-BP) (7). The whole data comprised the names, sequences and binding site motifs of 576 TFs, including the list of 8-mers evaluated in each PBM experiment with their corresponding E-score values and the resulting PWMs.

We relied on the following software: DSSP (version CMBI 2006) (43) and 3DNA (version 2.0) (44) provided the structural features of TFs and DNA, respectively; MATCHER, as distributed in the EMBOSS package (version 6.5.0) (45), was applied to refine alignments; BLAST (version 2.2.22) (46) was employed for homology detection; MODELLER (version 9.10) (47) was used to model-build TFs complexes with DNA (when necessary, DNA was modeled using 3DNA); and TM-align (version 20120126) (48) was utilized for structural superpositions.

Protein-DNA contact and interface

We defined a protein-DNA contact between an amino acid and a paired dinucleotide if the Cβ atom of the amino acid (Cα for glycines) was found within 15.0 Å from the geometric center of the dinucleotide, as delimited by the four phosphates of the two nucleotides and their associated partners in the complementary strand (9). Moreover, we defined a protein-DNA interface as follows. At the side of the protein, the interface was composed of all secondary structures (α-helices and β-sheets) containing at least one amino acid involved in a protein-DNA contact, including any intermediate regions. At the side of DNA, the interface was delimited by the first and last base pairs in dinucleotides contacted by amino acids. In a more restrictive definition, the DNA interface was delimited by the first and last base pairs in dinucleotides involving a minimum of 5 different protein-DNA contacts and at least one of them within 12.0 Å. This was defined as the "core" DNA interface.

Positive and negative 8-mers

Given a TF motif, its 8-mers were classified according to their ability to be bound by the TF (*i.e.* positives) or not (*i.e.* negatives). Specifically, 8-mers were considered positive (*i.e.* P8Ms) if their E-score ≥0.45 whereas negative 8-mers (*i.e.* N8Ms) had E-scores ≤0.37. Moreover, since PBMs don't specify whether an 8-mer is being recognized through the forward or reverse strand, we classified both strands of an 8-mer as either positive or negative. Unclassified 8-mers were considered dubious and discarded. Overall, the P8M/N8M ratio was 1 to 339.

Set of known TF-DNA complex structures

We examined all TF-DNA complex structures retrieved from the PDB in order to remove promiscuous residues and inadequate complexes. First of all, each structure was inspected and isolated nucleotides and overlapped DNA chains (*i.e.* RMSD between their backbones was <1.0 Å) were deleted. Then, complexes were analyzed using DSSP and 3DNA. The analysis led to the removal of residues that could not be recognized by any of these programs. Moreover, complexes that did not contain a double-stranded DNA molecule of at least 8 consecutive base pairs, according to 3DNA, were discarded. Finally, TF chains with DNA interfaces smaller than 8 base pairs were not considered. The remaining 583 TF-DNA complexes comprised a total of 1,341 TF chains, which constituted our set of known structures.

Dimer set

We identified all dimers in the set of known structures by grouping any two TF chains from the same PDB that: 1) belonged to the same structural family (*i.e.* both chains superimposed with a TM-score of at least 0.6 (49)); 2) had overlapping protein-DNA interfaces; and 3) had more than 5 residue-residue contacts between them as to form a binary complex (50).

Set of modeled TF-DNA complex structures

We modeled the TF interactions with different P8Ms as observed in the PBM data. The modeling of the whole complex was done separately for each TF, splitting the process in modeling the TF, the 8-mers and the complex.

Modeling of the TF

Given a TF, different structural models of the TF in complex with DNA were obtained as follows (see Figure 3): In step 1, the TF sequence was scanned for putative homologs in the set of known structures using BLAST. In step 2, we only used as templates those BLAST hits ensuring an alignment with enough percentage of sequence identity (above the twilight-zone curve (51)) and without gaps in the interface region. In step 3, the TF was realigned to the template sequences using MATCHER. In step 4, each alignment was used to create an optimized structural model of the TF using MODELLER. This procedure yielded several models of the same TF.

In addition, for TFs of the bHLH and bZIP families, since they recognize DNA as homo- or heterodimers, for each selected hit the dimer was modeled as follows: First, if the hit was already a homodimer, we used it as template. Otherwise, we searched the closest structural dimer to the hit in the set of dimers using TM-align and used the found dimer as new template. Then, the TF was realigned to both template chains (*i.e.* step 3) in order to generate a homodimer using MODELLER (*i.e.* step 4).

Modeling of the P8Ms

We used 3DNA to model the structure of the different P8Ms associated to the TF. We used as templates the DNA structures of the different hits obtained during the modeling of the TF in step 2 (see above). However, in order to model the interaction with the TF we needed to know the exact location of the TF-P8M interface. To place the interface between the TF and each P8M correctly, we realigned all P8Ms to the DNA sequence of the hit as follows: First, we constructed a multiple sequence alignment (MSA) around the most dominant P8M of the TF according to the PBM (*i.e.* the P8M with the highest E-score). Using that P8M as seed, the other P8Ms of the TF were incorporated to the MSA if: 1) they were single-nucleotide variants of seed P8M; or 2) they included a continuous gap of a maximum of 2 nucleotides at 3' or 5' side and no mismatches with the seed P8M. Second, if the sequence of the core DNA interface of the hit was found among the P8Ms of the MSA, the

P8Ms of the MSA could be realigned to it. Otherwise, the modeling using this hit as template was aborted.

Constructing the TF-P8M complexes

All TF-P8M complexes were modeled by superposing the models of the P8M and the model of TF to their corresponding templates using TM-align. Finally, the TF-P8M complex structure was optimized using MODELLER. In total, we obtained 52,419 TF-P8M complex models.

Non-redundant sets

We constructed two nr sets of TF-DNA complexes, named nr35 and nr70. Both sets contained known as well as modeled TF-DNA complex structures. Redundancy was removed at the level of the protein-DNA interface so that any two structures in sets nr35 and nr70 could not share more than 35 and 70% of protein-DNA contacts, respectively. The total number of known structures and models in each nr set were: 352 and 112 for nr35, and 767 and 1977 for nr70.

Statistical potentials implemented in SHAITAN

In a recent work, we derived 5 different statistical potentials to predict protein-DNA interactions (9). Among them, E_{S3DC} showed comparable results to other state-of-the-art methods when applied to predict the PWMs of 71 targets from the DREAM5 Challenge (14). Given a protein-DNA complex structure, the statistical potential E_{S3DC} was obtained by summing the potential of mean force PMF_{S3DC} over each protein-DNA contact, as defined by the amino acid *a* and the dinucleotide *mn*:

$$E_{S3DC} = \sum_{a,mn} PMF_{S3DC}(a,mn)$$

$$PMF_{S3DC}(a,mn) = -k_B T log\left(\frac{P\left(a,mn\middle|d_{a,mn},\theta_a,\theta_{mn}\right)P(\theta_a,\theta_{mn})}{P(a,mn|\theta_a,\theta_{mn})P\left(\theta_a\theta_{mn}\middle|d_{a,mn}\right)}\right)$$

Where k_B denotes the Boltzmann constant and T the standard temperature (300 K), $d_{a,mn}$ represents the contact distance between the amino acid *a* and the dinucleotide *mn*, and θ_a and θ_{mn} are their respective environments. The environment of an amino acid was defined by its hydrophobicity, degree of exposure and surrounding secondary structure, while the environment of a dinucleotide was specified by its constituting bases and three features regarding the interaction between the amino acid and the dinucleotide: 1) the closest strand to the amino acid; 2) the DNA-groove where the amino acid was located; and 3) the closest chemical group of the dinucleotide to the amino acid (9). Terms *P(*)* denote the probabilities of observing contacts with different characteristics. For example, $P(a,mn|d_{a,mn},\theta_a,\theta_{mn})$ is the conditional probability of observing the amino acid *a* and the dinucleotide *mn*, in their respective environments θ_a and θ_{mn}, within a distance $d_{a,mn}$.

In this work, we further decomposed $\mathrm{PMF}_{\mathrm{S3DC}}$ in three different terms:

$$PMF_{S3DC}(a,mn) =$$

$$k_B T log\left(\frac{P(\theta_a\theta_{mn}|d_{a,mn})}{P(\theta_a,\theta_{mn})}\right) \qquad 1$$

$$- k_B T log\left(P(a,mn|d_{a,mn},\theta_a,\theta_{mn})\right) \qquad 2$$

$$+ k_B T log\left(P(a,mn|\theta_a,\theta_{mn})\right) \qquad 3$$

Where term *1* is $\mathrm{PMF}_{\mathrm{3DC}}$ (see (9)), and terms *2-3* are two new potentials. One of them is dependent of the distance at which a contact is observed, while the other one is distance independent. These potentials were named as $\mathrm{PMF}_{\mathrm{S3DC}}$ distance-dependent (*i.e.* $\mathrm{PMF}_{\mathrm{S3DCdd}}$) and $\mathrm{PMF}_{\mathrm{S3DC}}$ distance-independent (*i.e.* $\mathrm{PMF}_{\mathrm{S3DCdi}}$). As for $\mathrm{E}_{\mathrm{S3DC}}$, statistical potentials $\mathrm{E}_{\mathrm{S3DCdd}}$ and $\mathrm{E}_{\mathrm{S3DCdi}}$ were the result of summing their corresponding PMFs over all protein-DNA contacts *a,mn*.

Probabilities *P(*)* were derived from the frequencies of observed contacts in the nr sets. Still, some protein-DNA contacts might not be covered. In such cases, their PMFs were estimated from contacts involving the same dinucleotide but different amino acids. For example, if the probability of a protein-DNA contact between the amino acid *a* and the dinucleotide *mn* was unknown, but we knew the probabilities of similar contacts, let them be specified by *b,mn*, then the probability of the contact *a,mn* was calculated as follows:

$$P(a,mn) = \sum_{b \neq a} P(b \to a)\, P(b,mn)$$

Where *P(b→a)* is the BLOSUM62 transition probability (52) from amino acid *b* to *a*. Therefore, by introducing the previous equation, the PMF for contact *a,mn* was approximated as:

$$PMF(a,mn) = -k_B T log\left(\sum_{b \neq a} P(b \to a)\, e^{-PMF(b,mn)/k_B T}\right)$$

Moreover, the sum inside the logarithm of last equation was approached by Taylor's (53), neglecting orders higher than 2, which resulted in:

$$PMF(a,mn) = -k_B T\left(log\big(P(b_{max} \to a)\big) - \frac{PMF(b_{max}, mn)}{k_B T} + \sum_{\substack{b \neq a \\ b \neq b_{max}}} \frac{P(b \to a) e^{-PMF(b,mn)/k_B T}}{P(b_{max} \to a) e^{-PMF(b_{max},mn)/k_B T}}\right)$$

Where b_{max} was the amino acid that maximized the following product:

$$P(b \to a) e^{-PMF(b,mn)/k_B T}$$

Additionally, we transformed PMFs into Z-scores as follows. Given a protein-DNA contact between the amino acid *a* and the dinucleotide *mn* (*i.e. a,mn*), the Z-score was calculated from the difference between the contact PMF (*i.e. PMF(a,mn)*) and the averaged PMFs of all contacts involving the same dinucleotide (*i.e. μ*) divided by the standard deviation (*i.e. σ*):

$$Z - score(a,mn) = \frac{PMF(a,mn) - \mu}{\sigma}$$

Finally, we derived two types of statistical potentials called general and family-specific potentials. On the one hand, general potentials were derived from nr35 complexes. On the other hand, given a TF-DNA complex, family-specific potentials were derived from a subset of nr70 belonging to the same structural family of the complex (see Dimer set).

Five-fold cross-validation

In order to evaluate the ability of SHAITAN to distinguish P8Ms from N8Ms, we conducted a five-fold cross-validation on nr35 and nr70. For each set, structure models were split into five folds. Four of them were used to derive the statistical potentials and models in the remaining fold were used for testing. This process was repeated five times, changing each time the tested fold. For each tested model, all possible 8-mers were evaluated by threading them with one of the statistical potentials implemented in SHAITAN (*i.e.* E_{S3DC}, E_{S3DCdd} or E_{S3DCdi}) and the resulting scores were scaled between ±0.5. We only kept the scores of the modeled P8M as well as of 500 randomly selected N8Ms.

We tested several score cut-offs. For E_{S3DC} and E_{S3DCdd}, a prediction under the cut-off was considered "positive" while over the cut-off it was considered "negative". For E_{S3DCdi} it was the opposite, as high scores were meant for correct predictions. True positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) were defined accordingly. We computed the TPR (also termed sensitivity or recall), the false-positive rate (FPR), and the PPV (also termed precision). The resulting AUROCs, AUPRs and PPVs were averaged from the five folds tested.

Prediction of DREAM5 PWMs

Given a DREAM5 TF-DNA complex structure, we used a sliding window of 8 nucleotides to score all possible 8-mers along the DNA-interface with family-specific statistical potentials. At each step, the resulting scores were scaled between ±0.5 and all 8-mers passing the cut-off of ±0.45 (sign depending on the potential) were used to calculate the probabilities of the different nucleotides in each PWM column, approached from the nucleotide frequency.

Prediction of homeodomain protein-PWMs

Given a target DNA sequence bound by a homeodomain, this was threaded in all TF-DNA complexes of the set of known structures classified as members of the homeobox Pfam family (54). We used the statistical potential E_{S3DCdd} to rank the threaded structures. The homeobox structure with the most negative score was further used to construct a PWM of the protein as follows: First, the sequences of all homeodomains with associated PBM data (26) were aligned to the homeobox structure using BLAST. We discarded all homeodomains without at least one P8M matching the core DNA interface of the homeobox. The remaining homeodomains were used to construct a MSA using Clustal

Omega (55). Next, the MSA was realigned to the sequence of the *Drosophila* Engrailed protein (56). Special attention was taken on the region comprising the 3 specificity residues that bind DNA (*i.e.* residues 47-54). Finally, the realigned MSA was used to calculate the PWM and the frequencies of amino acids at each position. For this test, we used family-specific potentials derived exclusively from homeobox structures and homeodomain PBM data (26).

ACKNOWLEDGEMENTS

O.F. and B.O. acknowledge the support of FEDER BIO2011-22568 grant from the Spanish Ministry of Science and Innovation (MICINN). M.T.W. was supported by fellowships from CIHR and the Canadian Institute for Advanced Research (CIFAR) Junior Fellows Genetic Networks Program. T.H.R. is a CIFAR scholar and was supported by grants from the Canadian Institutes of Health Research (MOP-490425 and MOP-459114).

REFERENCES

1. Bulyk ML. Computational prediction of transcription-factor binding site locations. Genome Biol. 2003 Dec 23;5(1):201.

2. Buck MJ, Lieb JD. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics. 2004 Mar;83(3):349–60.

3. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009 Oct;10(10):669–80.

4. Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc. 2009 Mar;4(3):393–411.

5. Bulyk ML. Discovering DNA regulatory elements with bacteria. Nat Biotechnol. 2005 Aug;23(8):942–4.

6. Reece-Hoyes JS, Diallo A, Lajoie B, Kent A, Shrestha S, Kadreppa S, et al. Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. Nat Methods. 2011 Dec;8(12):1059–64.

7. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell. 2014 Sep 11;158(6):1431–43.

8. Das MK, Dai H-K. A survey of DNA motif finding algorithms. BMC Bioinformatics. 2007 Nov 1;8(Suppl 7):S21.

9. Fornes O, Garcia-Garcia J, Bonet J, Oliva B. On the Use of Knowledge-Based Potentials for the Evaluation of Models of Protein-Protein, Protein-DNA, and Protein-RNA Interactions. Adv Protein Chem Struct Biol. 2014;94:77–120.

10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–42.

11. Luscombe NM, Thornton JM. Protein–DNA Interactions: Amino Acid Conservation and the Effects of Mutations on Binding Specificity. J Mol Biol. 2002 Jul 26;320(5):991–1009.

12. Benos PV, Bulyk ML, Stormo GD. Additivity in protein–DNA interactions: how good an approximation is it? Nucleic Acids Res. 2002 Oct 15;30(20):4442–51.

13. AlQuraishi M, McAdams HH. Three enhancements to the inference of statistical protein-DNA potentials. Proteins Struct Funct Bioinforma. 2013;81(3):426–42.

14. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, et al. Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol. 2013 Feb;31(2):126–34.

15. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning [Internet]. New York, NY, USA: ACM; 2006 [cited 2014 Sep 1]. p. 233–40. Available from: http://doi.acm.org/10.1145/1143844.1143874

16. Wolfe SA, Nekludova L, Pabo CO. DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct. 2000;29(1):183–212.

17. Pavletich NP, Pabo CO. Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. Science. 1993 Sep 24;261(5129):1701–7.

18. Segal DJ, Crotty JW, Bhakta MS, Barbas III CF, Horton NC. Structure of Aart, a Designed Six-finger Zinc Finger Peptide, Bound to DNA. J Mol Biol. 2006 Oct 20;363(2):405–21.

19. Contreras-Moreira B. 3D-footprint: a database for the structural analysis of protein–DNA complexes. Nucleic Acids Res. 2010 Jan 1;38(suppl 1):D91–7.

20. Gabdoulline R, Eckweiler D, Kel A, Stegmaier P. 3DTF: a web server for predicting transcription factor PWMs using 3D structure-based energy calculations. Nucleic Acids Res. 2012 Jul 1;40(W1):W180–5.

21. Lin C-K, Chen C-Y. PiDNA: predicting protein–DNA interactions with structural models. Nucleic Acids Res. 2013 Jul 1;41(W1):W523–30.

22. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007 Feb 26;8(2):R24.

23. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 2009 Jul 1;37(suppl 2):W202–8.

24. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011 Apr 1;27(7):1017–8.

25. Ledneva RK, Alekseevskiĭ AV, Vasil'ev SA, Spirin SA, Kariagina AS. Structural aspects of homeodomain interactions with DNA. Mol Biol (Mosk). 2001 Oct;35(5):764–77.

26. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, et al. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. Cell. 2008 Jun 27;133(7):1266–76.

27. Fu F, Voytas DF. Zinc Finger Database (ZiFDB) v2.0: a comprehensive database of C2H2 zinc fingers and engineered zinc finger arrays. Nucleic Acids Res. 2013 Jan 1;41(D1):D452–5.

28. Walhout AJM. Gene-centered regulatory network mapping. Methods Cell Biol. 2011;106:271–88.

29. Turner D, Kim R, Guo J. TFinDit: transcription factor-DNA interaction data depository. BMC Bioinformatics. 2012 Sep 3;13(1):220.

30. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 2006 Nov;24(11):1429–35.

31. Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, Tsui K, et al. A Library of Yeast Transcription Factor Motifs Reveals a Widespread Function for Rsc3 in Targeting Nucleosome Exclusion at Promoters. Mol Cell. 2008 Dec 26;32(6):878–87.

32. Silva EKD, Gehrke AR, Olszewski K, León I, Chahal JS, Bulyk ML, et al. Specific DNA-binding by Apicomplexan AP2 transcription factors. Proc Natl Acad Sci. 2008 Jun 17;105(24):8393–8.

33. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, et al. Diversity and Complexity in DNA Recognition by Transcription Factors. Science. 2009 Jun 26;324(5935):1720–3.

34. Grove CA, De Masi F, Barrasa MI, Newburger DE, Alkema MJ, Bulyk ML, et al. A Multiparameter Network Reveals Extensive Divergence between C. elegans bHLH Transcription Factors. Cell. 2009 Jul 23;138(2):314–27.

35. Lesch BJ, Gehrke AR, Bulyk ML, Bargmann CI. Transcriptional regulation and stabilization of left–right neuronal identity in C. elegans. Genes Dev. 2009 Feb 1;23(3):345–58.

36. Scharer CD, McCabe CD, Ali-Seyed M, Berger MF, Bulyk ML, Moreno CS. Genome-Wide Promoter Analysis of the SOX4 Transcriptional Network in Prostate Cancer Cells. Cancer Res. 2009 Jan 15;69(2):709–17.

37. Zhu C, Byers KJRP, McCord RP, Shi Z, Berger MF, Newburger DE, et al. High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res. 2009 Apr 1;19(4):556–66.

38. Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinás M. Identification and Genome-Wide Prediction of DNA Binding Specificities for the ApiAP2 Family of Regulators from the Malaria Parasite. PLoS Pathog. 2010 Oct 28;6(10):e1001165.

39. Wei G-H, Badis G, Berger MF, Kivioja T, Palin K, Enge M, et al. Genome‐wide analysis of ETS‐family DNA‐binding in vitro and in vivo. EMBO J. 2010 Jul 7;29(13):2147–60.

40. Chang KN, Zhong S, Weirauch MT, Hon G, Pelizzola M, Li H, et al. Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. eLife. 2013 Jun 11;2:e00675.

41. Grau J, Posch S, Grosse I, Keilwagen J. A general approach for discriminative de novo motif discovery from high-throughput data. Nucleic Acids Res. 2013 Sep 20;gkt831.

42. Sebé-Pedrós A, Ariza-Cosano A, Weirauch MT, Leininger S, Yang A, Torruella G, et al. Early evolution of the T-box transcription factor family. Proc Natl Acad Sci. 2013 Oct 1;110(40):16050–5.

43. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–637.

44. Lu X-J, Olson WK. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc. 2008 Jul;3(7):1213–27.

45. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 2000 Jun 1;16(6):276–7.

46. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389–402.

47. Eswar N, Webb B, Marti-Renom MA, Madhusudhan M s., Eramian D, Shen M, et al. Comparative Protein Structure Modeling Using Modeller. Current Protocols in Bioinformatics [Internet]. John Wiley & Sons, Inc.; 2006 [cited 2013 Sep 24]. Available from: http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi0506s15/abstract

48. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005 Jan 1;33(7):2302–9.

49. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics. 2010 Apr 1;26(7):889–95.

50. Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. Nat Methods. 2013 Jan;10(1):47–53.

51. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999 Feb 1;12(2):85–94.

52. Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol. 2004 Aug;22(8):1035–6.

53. Odibat ZM, Shawagfeh NT. Generalized Taylor's formula. Appl Math Comput. 2007 Mar 1;186(1):286–93.

54. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2011 Nov 29;40(D1):D290–301.

55. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high‐quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Jan 1;7(1):539.

56. Kissinger CR, Liu BS, Martin-Blanco E, Kornberg TB, Pabo CO. Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: a framework for understanding homeodomain-DNA interactions. Cell. 1990 Nov 2;63(3):579–90.

57. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—A visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605–12.

58. Glover JNM, Harrison SC. Crystal structure of the heterodimeric bZIP transcription factor c-Fos–c-Jun bound to DNA. Nature. 1995 Jan 19;373(6511):257–61.

59. Fraenkel E, Pabo CO. Comparison of X-ray and NMR structures for the Antennapedia homeodomain–DNA complex. Nat Struct Mol Biol. 1998 Aug;5(8):692–7.

60. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: A Sequence Logo Generator. Genome Res. 2004 Jun 1;14(6):1188–90.

Table 1. SHAITAN AUROCs and AUPRs using statistical potentials E_{S3DC}, E_{S3DCdd} and E_{S3DCdi}.

		E_{S3DC}		E_{S3DCdd}		E_{S3DCdi}	
Complexes	Potentials	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
PDB	general	0.708	0.005	0.732	0.005	0.656	0.004
PDB+PBM	general	0.781	0.008	0.804	0.010	0.724	0.006
PDB+PBM	family	0.920	0.040	0.960	0.123	0.912	0.057

AUPR expected by a random distribution is 0.002.

Columns 1-2 show if SHAITAN statistical potentials were derived from PDB structures alone (*i.e.* PDB) or in combination with PBM models (*i.e.* PDB+PBM), and if they were general or family-specific. Columns 3-8 display averaged AUROCs and AUPRs from using a five-fold cross-validation approach for SHAITAN using statistical potentials E_{S3DC}, E_{S3DCdd} and E_{S3DCdi}.

Table 2. SHAITAN AUROCs and AUPRs of family-specific potentials E_{S3DC}, E_{S3DCdd} and E_{S3DCdi} for different TF families.

	E_{S3DC}		E_{S3DCdd}		E_{S3DCdi}	
Family	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
AFT	1	0	0.99	0.083	0.043	0.001
AP2	0.997	0.196	1	0.431	0.995	0.102
ARID/BRIGHT						
bHLH	0.91	0.025	0.932	0.156	0.913	0.071
bZIP	0.984	0.142	0.993	0.148	0.81	0.02
C2H2 zinc finger	0.924	0.059	0.938	0.056	0.862	0.02
E2F						
Ets	0.943	0.062	0.975	0.208	0.939	0.087
Forkhead	0.945	0.055	0.974	0.143	0.938	0.068
GATA	0.997	0.468	0.993	0.157	0.814	0.008
Homeodomain	0.93	0.04	0.959	0.113	0.932	0.064
IRF	0.921	0.057	0.952	0.048	0.885	0.011
MADS box						
Myb/SANT	0.962	0.034	0.936	0.073	0.646	0.003
NAC/NAM	0.971	0.061	0.987	0.111	0.921	0.04
Nuclear receptor	0.88	0.023	0.958	0.149	0.9	0.062
Paired box	0.934	0.02	0.989	0.121	0.99	0.156
POU	0.99	0.448	0.994	0.586	0.968	0.129
SMAD	0.972	0.033	0.928	0.014	0.094	0.001
Sox	0.996	0.541	0.991	0.424	0.905	0.065
T box	0.905	0.055	0.946	0.072	0.832	0.026
TBP	0.999	0.54	0.999	0.565	0.994	0.173
WRKY						
Zinc cluster	1	0.5	0.999	0.208	0.739	0.004

AUPR expected by a random distribution is 0.002.

Column 1 shows the different TF families studied. Columns 2-7 display averaged AUROCs and AUPRs obtained from a five-fold cross-validation approach for SHAITAN using family-specific statistical potentials E_{S3DC}, E_{S3DCdd} and E_{S3DCdi}. Empty columns indicate no members of that family could be tested.

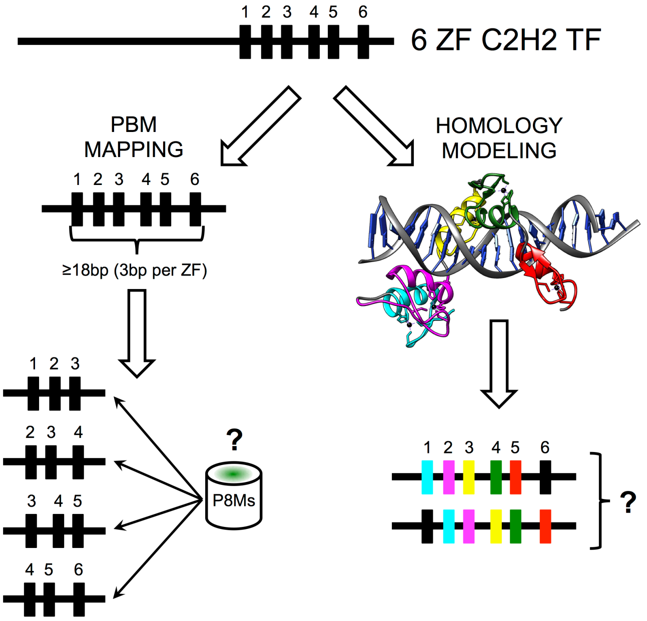


Figure 1. Problems related to the use of PBM data from members of the C2H2 zinc finger family. We provide a 6-finger protein as an example. On the one hand, each finger is able to recognize 3 consecutive base pairs (bps), which makes the binding site of the protein at least 18 bps long. Therefore, when mapping the PBM data of the protein, we cannot know which combination of fingers interacted with the probes of the PBM experiment. On the other hand, when modeling the protein using a 5-finger template (17), we cannot know which of the possible models contains the interacting fingers of the PBM experiment. The structure image was created with the UCSF Chimera package (57).

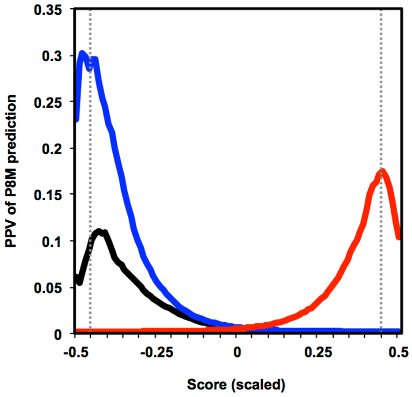


Figure 2. SHAITAN averaged PPVs obtained from a five-fold cross-validation approach using family-specific statistical potentials E_{S3DC} (black), E_{S3DCdd} (blue) and E_{S3DCdi} (red). Dashed grey lines are used to indicate the highest PPV peaks for the different statistical potentials (approximately at ±0.45 score cut-off depending on the potential).

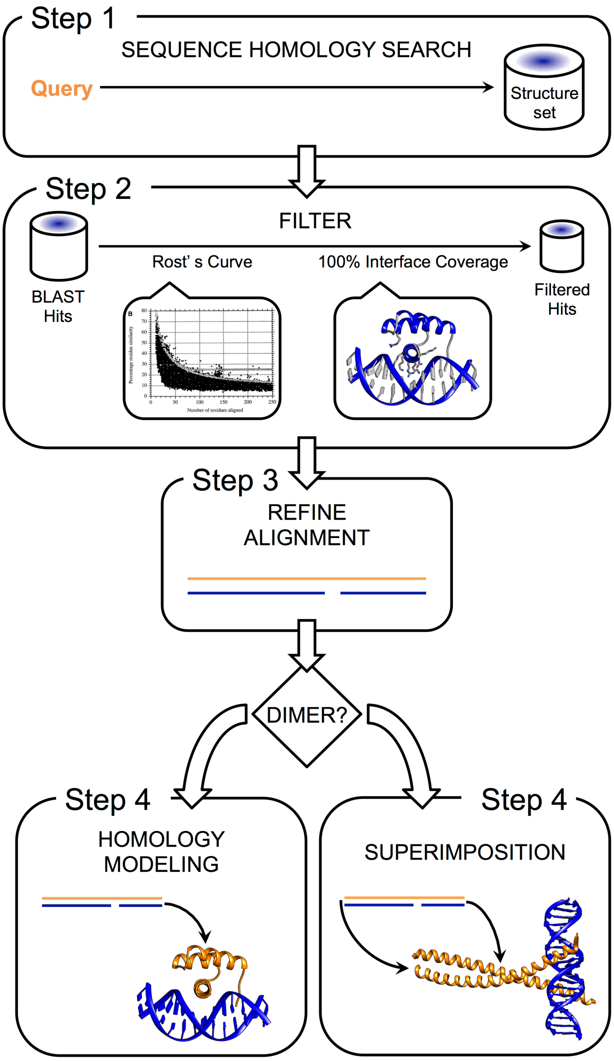


Figure 3. Homology modeling of TF-DNA complexes. Step 1: sequence homology search. Step 2: filter results from step 1 by sequence identity and protein-DNA interface coverage. Step 3: alignment optimization. Step 4: model building of the TF three-dimensional structure. If the TF works as a dimer, superimpose the alignment from step3 to the closest structural dimer and model the TF homodimer. Structural images were created with the UCSF Chimera package (57–59).

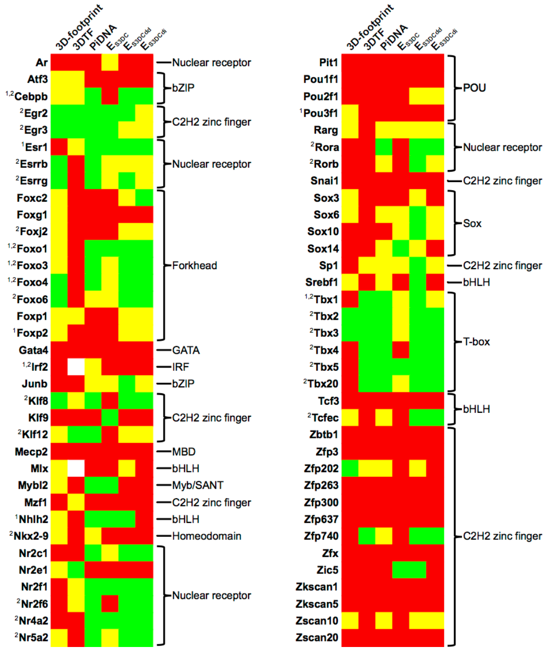
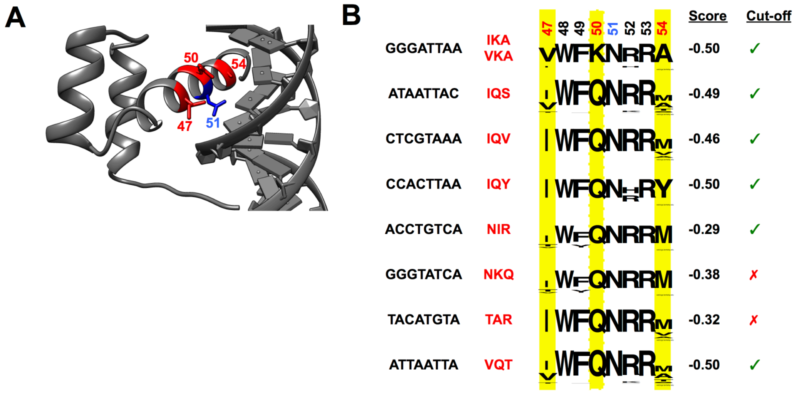


Figure 4. Comparison between the PWMs predicted by 3D-footprint, 3DTF, PiDNA and SHAITAN, using statistical E_{S3DC}, E_{S3DCdd}, and E_{S3DCdi}, and the experimentally determined PWMs for 70 DREAM5 targets. The names and families of the different targets are shown at both sides of the predictions. Superindices ^{1} and ^{2} indicate if the prediction was performed using a real PDB structure (*i.e.* no modeling was required) and if the target PWM was easy to predict (*i.e.* the target PWM bore some resemblance with the nucleotide sequence of the TF-DNA complex, according to FIMO). Finally, green, yellow, and red squares indicate if the predictions of the different methods were good, mediocre or bad, respectively.



		Score	Cut-off
GGGATTAA	IKA VKA	-0.50	✓
ATAATTAC	IQS	-0.49	✓
CTCGTAAA	IQV	-0.46	✓
CCACTTAA	IQY	-0.50	✓
ACCTGTCA	NIR	-0.29	✓
GGGTATCA	NKQ	-0.38	✗
TACATGTA	TAR	-0.32	✗
ATTAATTA	VQT	-0.50	✓

Figure 5. Comparison between homeodomain specificity residues and SHAITAN predicted logos for determined DNA sequences. A) Protein-DNA complex structure of the *Drosophila* Engrailed protein (56). The three primary specificity amino acids discussed in the text are shown in red. Also, the highly conserved Asn-51 across the homeodomain family (25) is shown in blue. B) Logos comprising residues 47-54 of homeodomain proteins predicted to interact with each of the displayed DNA sequences. The positions of the three specificity residues are highlighted in yellow. Additionally, the different combinations of specificity amino acids known to interact with each DNA sequence are shown in red (26). "Score" and "Cut-off" columns indicate the SHAITAN score of the prediction and whether if it passed the energy cut-off of -0.45 or not (*i.e.* a green tick or a red cross, respectively). Engrailed structural image was created with the UCSF Chimera package (57) and logos were obtained using WebLogo (60).

4. Discussion

“I think that in the discussion of natural problems we ought to begin not with the Scriptures, but with experiments, and demonstrations.”

-Galileo Galilei

Eukaryotic genomes encode for thousands of different proteins and RNAs (262). Some of these genes are constitutively expressed, while others are expressed in a tightly controlled manner in only part of the organism, or under particular conditions as for example during development or disease. To understand how differential gene expression is regulated at a genome-wide or systems level, it is key to identify all TFs and their binding sites in promoter or enhancer regions of genes, and how and when they interact to affect gene expression. In this regard, the majority of efforts have focused on identifying the binding preferences of TFs. However, we are still far from having the whole picture of gene regulatory networks (263).

During the past decade, the appearance of several high-throughput techniques has allowed the characterization of TF-binding sites at large-scale (revised in section 1.3.3), which has supposed a major breakthrough on the field. Still, the application of such protocols is both laborious and difficult and as a result, only a small fraction of eukaryotic TFs has been profiled (201). Thus, the development of computational approaches for rapid and accurate mapping of TFs and their binding sites along the genome is foremost, especially as a complement to current experimental procedures.

My thesis has focused on the study of PDIs from a bioinformatics perspective. Specifically, the work I have developed during my doctorate can be divided into:

1) Study redundancy in gene regulatory networks due to TFs
2) Predict DNA sequences that can be bound by a TF and *vice versa*

From this point forward, I will proceed with the discussion of this thesis by explaining the contribution of my research to the field and finally, I will devise new strategies and consider future perspectives with a special focus on functional genomics and the analysis of gene regulation.

4.1 Redundancy between transcription factors: the role of protein-protein interactions

Even though the knowledge on PPIs is far from complete, PPI data has been broadly used to transfer annotation between proteins based on the interologs approach (264,265). This is, two proteins are more likely to be homologous if they also have common PPIs. As revised in section 1.2.3, in eukaryotes, it is by now fairly clear that TFs act cooperatively forming enhanceosomes. Upon assembly, these TF-complexes are able to bring RNA polymerase to the promoter and activate gene transcription. Genome-wide analysis of the binding preferences for members of two different TF families have revealed that binding patterns are conserved between homologous TFs (258,266). Therefore, a pair of homologous TFs that in addition interacted with the same proteins should be able, in theory, to promote the transcription of the same genes.

I further explored this hypothesis in collaboration with Dr. Anthony Gitter and Prof. Ziv Bar-Joseph from the Carnegie Mellon University (USA) (see section 3.1.1). In a previous work, Hu *et al.* (267) had knocked out 269 budding yeast TFs, one at a time, and had compared the differential expression of the different genes with the binding sites that had been previously generated for 188 of these TFs (268). Surprisingly, they observed that only 3% of the genes had been affected by the knockout and similarly, only 3% of the affected genes had been bound by the knocked out TF. To further explain these results, I did the following:

1) Find pairs of homologous TFs (using BLAST (269) *e*-values)
2) Group the TFs into different Pfam (270) families (by DBDs)
3) Calculate the percentage of shared PPIs between TFs
4) Automate remote homology detection using previous steps

For TFs with a close homolog there was no overlap between their binding and knockout data. In contrast, TFs without any assigned homolog (using a BLAST *e*-value of 10^{-3}) had an overlap of more than 12%. Moreover, the use of Pfam family classification as a measure of similarity resulted in comparable results than when using BLAST homology: for TFs classified in larger groups the overlap was lower than for those classified in smaller groups. A similar trend was observed when using PPIs: as the percentage of shared PPIs between TFs increased, the overlap decreased. Specifically, for TFs without homologs and sharing no PPI data the overlap was greater than 13%.

Certain pairs of homologous TFs identified in this work bind to an overlapping group of target genes. Hence, it is not surprising than their knockouts had a small effect of their targets expression. For example, Met31 and Met32 have a large overlapping set of target genes (>60%), and none of them was differentially expressed after the knockouts. Another example was Fkh1 and Fkh2. These TFs bind to a partially overlapped set of genes. However, the binding of Fkh1 to Fkh2 targets is enhanced in the absence of Fkh2 and *vice versa* (271), suggesting that compensation might occur beyond their common targets, as predicted.

This kind of compensation may happen due to competition between the two TFs that resolves in the absence of one of them. Another possibility is the activity of one TF may be enhanced in the absence of its homolog (272). The people of the Systems Biology Group in the Carnegie Mellon University further inspected this idea by looking at the expression levels of the closest TF homologs predicted to be compensating for the knockout. As expected, they could not find any TF whose expression level had been significantly decreased. Still, a significant increase was observed for a few cases, which might be driven by PPIs.

The findings obtained in this work strongly suggest the knockouts might be compensated by homologs sharing many PPI with the deleted TF, thus agreeing with our initial hypothesis. The approach was further automated and generalized to infer the fold of remote homologs, and is included in ModLink+ (see section 3.1.2).

4.2 Statistical potentials for protein-DNA interactions: how good are they?

TF-binding specificities represent the cornerstone on functional genomics and the analysis of gene regulation. However, only about 1% of eukaryotic TFs have been profiled using “omics” techniques (201). In this regard, the development of computational tools to complement experimental procedures is fundamental (see section 1.3.5). One successful approach to predict the binding sites of a TF is the structural analysis of its complex with DNA using statistical potentials (see section 3.2.1). Briefly, given a TF-DNA complex structure, different DNA sequences are threaded and the binding energies of the resulting interactions are calculated. Best scoring sequences are then considered to be bound by the TF.

However, their application still has some limitations. One of them is the small number of TF-DNA complex structures available in the PDB. Statistical potentials are usually derived from a non-redundant set of PDB structures. This redundancy is typically removed on the TF side of the complex. Yet, TFs can recognize different binding sites and in addition, members of the same family of TFs can bind to distinct DNA sequences (211), which can result in statistical potentials affected by both low-count and low diversity of binding patterns. Moreover, statistical potentials are derived under the wrong assumption the contribution of the different DNA base pairs to the binding energy of the complex is independent from each other (261). In sections 3.2.1 and 3.3.1 I tackled both problems by:

1) Develop statistical potentials for contacts between amino acids and dinucleotides (*i.e.* pairs of consecutive nucleotides along the DNA sequence)

2) Augment the coverage of the potentials by including PBM data

Both improvements are included in an approach named SHAITAN, which stands for Statistical-potentials using a Homology-based Approach to predict the Interactions between Transcription factors And Nucleotide sequences (see section 3.3.1). SHAITAN is better than other current state-of-art methods available online on the prediction of TF-PWMs. Moreover, it can be applied to detect the best TF able to bind a particular DNA sequence. Furthermore, thanks to an automated pipeline for modeling TF-DNA complexes, SHAITAN is applicable to annotate TF-binding sites at large-scale, thus providing the scientific community with a powerful tool that could be used to fill the existing gaps in gene regulatory networks.

One of the most important things when benchmarking SHAITAN was to define its limits in terms of reliable predictions. I conducted a five-fold cross-validation on two non-redundant sets of TF-DNA complexes in order to evaluate SHAITAN's ability to distinguish correct from incorrect complexes as defined by sequences known to be bound or not by each TF. The results from the evaluation allowed me to select the best cut-off in terms of positive predictive value. However, this has not been done in other cases. As examples I illustrate two other methods against which SHAITAN was compared (see section 3.3.1): 3DTF (223) and PiDNA (226).

For example, 3DTF was benchmarked using two different analyses on TFs having both known binding sites in TRANSFAC (163) or UniProbe (240) and a PDB structure of its complex with DNA. On the one hand, for each TF-DNA complex, they compared the energies calculated for random sequences with the energies from the known binding sequences. As expected, binding site sequences were assigned lower energies than random ones. On the other hand, they compared the produced PWMs for each TF-DNA complex to the TRANSFAC and UniProbe PWMs. They reported 3DTF-PWMs were more similar to the PWMs of the corresponding TF than to other PWMs. The other example, PiDNA, was evaluated using three different sets in order to: 1) compare its generated PWMs against real PWMs; 2) identify binding sites with high specificity; and 3) predict known binding sites from random sequences. However, none of them removed redundancy with respect to the evaluation tests and, as a result, there is no way to know their performance. Moreover, on the prediction of known binding sites from random sequences using PiDNA, authors used the area under the receiver operating characteristic curve (AUC) as a measure of performance. However, the AUC was not informative because they were dealing with very unbalanced data (they tested 10,000 random sequences for each known binding site), so that even a high AUC score could imply a high number of false positives (273).

4.3 Future challenges I: designing *de novo* transcription factors

The engineering of TFs, especially for zinc finger proteins (ZFPs), provides an important tool for studying gene regulation and modify the genome (*i.e.* genome editing) because they can be used to target virtually any desired DNA sequence (274).

As shown in section 3.3.1, SHAITAN can be applied to predict TFs able to bind a DNA sequence of interest by means of a PWM. However, such approach is limited by the knowledge on existing PBM data. Alternatively, given a TF-DNA complex structure, the prediction of the TF sequence can be done by approximating the probability of each amino acid from the energies of the different protein-DNA contacts. As shown in section 3.3.1, the energy of a TF-DNA complex structure is calculated by summing the potentials of mean force (PMF) over each protein-DNA contact, as defined by the amino acid *a* and the dinucleotide *mn* (*i.e. a,mn*):

$$E = \sum_{a,mn} PMF(a,mn)$$

This equation can be further decomposed to obtain the contribution to the energy of each individual amino acid:

$$E = \sum_{i} E_a^i$$

Where E_a^i is the energy of the amino acid a (which can be any of the 20 proteinogenic amino acids) at position i. In SHAITAN, E_a^i is independent of all other amino acids. Therefore, a PWM can be calculated from the probabilities of amino acids a at different positions *i* approximated using the Boltzmann formula:

$$P_a^i = \frac{e^{-E_a^i}}{\sum_b e^{-E_a^i}}$$

Where P_a^i is the probability of amino acid a, and *b* can be any of the 20 proteinogenic amino acids. The result is a PWM of the TF sequence of the evaluated structure that accommodates for different amino acids at determined positions (*i.e.* the positions where the protein contacts the DNA). Although this approach performs worse than the one shown for homeodomains (see section 3.3.1), it has the advantage that it can be applied even when there is no available PBM data.

Then, using the PWM, the most probable TF sequences could be derived and tested for being able to fold as the original structure or not using energy functions for protein folding such as ProSA (275), thus filtering those sequences that would not fold. The generated TFs would be finally validated experimentally in order to ascertain their ability to bind the DNA sequence of interest.

Still, in order to make SHAITAN applicable to genome editing, by engineering zinc finger endonucleases, the performance on this family should be improved by deriving statistical potentials using information on individual zinc fingers (247) instead of PBM data.

4.4 Future challenges II: towards modeling gene regulatory networks

A gene regulatory network (GRN) is a collection of DNA regions in a cell that interact with each other indirectly through their RNAs and protein expression products, thereby governing the cell's gene expression. GRNs are typically described as network models where gene dependencies are depicted by a directed graph, whose nodes represents genes and edges lead from a regulator (often a TF) to its targets. The edges preferably are used to represent dependencies at the transcriptional level, which are mediated by a TF binding to a regulatory region near the promoter region of the target gene. Thus, three different types of information are required to generate a GRN:

1) The spatio-temporal expression pattern of the TFs
2) The information on the TF-binding sites
3) A causal link between the TF activity and the expression of the target gene

Given the laborious task of identifying all nodes and edges in a GRN, most networks are at a small to medium size (276–282). Even in yeast, the most well studied eukaryote, the available information is far from complete. This is because many causal conditions still remain to be tested, and because for many TFs their low specificity binding sites may be difficult to analyze. For higher eukaryotes the situation is worse, in part due to the fact that less than 1% of eukaryotic TFs have been profiled (201). In addition, in higher eukaryotes, TFs act cooperatively forming enhanceosomes, with some of its members being located far from the gene promoter (*e.g.* enhancers). In such cases, it may be difficult to infer the gene that is being regulated by the enhanceosome.

In this context, information on chromatin interaction maps (283) and TF occupancy on nucleosome free regions (178–180) could be exploited together with TF-binding site data to predict putative enhanceosomes. A very simple approach using SHAITAN could be as follows:

For each interaction involving two or more DNA regions…

1) Identify the most likeable TFs able to bind to each region
2) Filter any TFs that could not be connected to any of the TFs assigned in the remaining regions using PPIs (*e.g.* at distance 2)

The resulting putative enhanceosomes could be further filtered by using microarray expression data specific for the tissue of interest (284).

5. Conclusions

“I know one thing: that I know nothing.”
-Socrates

This section describes in short the main achievements of the work presented in this thesis.

i. Redundancy in gene regulatory networks has been explained by homologous TFs sharing protein-protein interactions.

ii. The previous conclusion has been generalized and integrated in ModLink+, an online and user-friendly tool to infer the fold of remote homologs.

iii. Different applications of statistical potentials have been reviewed: the assessment of structural models, the ranking of docking possess, the detection of interfaces and the prediction of protein-DNA interactions.

iv. Split-statistical potentials for protein-DNA interactions have been proposed and vindicated.

v. Protein binding microarray data has been shown to improve the performance of statistical potentials for predicting transcription factor-DNA association.

vi. An automated modeling pipeline has been proposed to create protein-DNA complex structures.

vii. A homology-based approach to predict transcription factor-DNA interactions, SHAITAN, has been developed. SHAITAN has been demonstrated to be superior to current state-of-art methods for the prediction of transcription factor binding sites and to predict TFs able to bind a particular DNA sequence.

6. Appendix

6.1 Networks of protein-protein interactions: from uncertainty to molecular details

My contribution to this manuscript was to review methods to study co-expressed and co-localized proteins as well as the application of PPI for protein functional annotation.

Abstract

Proteins are the bricks and mortar of cells. The work of proteins is structural and functional, as they are the principal element of the organization of the cell architecture, but they also play a relevant role in its metabolism and regulation. To perform all these functions, proteins need to interact with each other and with other bio-molecules, either to form complexes or to recognize precise targets of their action. For instance, a particular transcription factor may activate one gene or another depending on its interactions with other proteins and not only with DNA. Hence, the ability of a protein to interact with other biomolecules, and the partners they have at each particular time and location can be crucial to characterize the role of a protein. Proteins rarely act alone; they rather constitute a mingled network of physical interactions or other types of relationships (such as metabolic and regulatory) or signaling cascades. In this context, understanding the function of a protein implies to recognize the members of its neighborhood and to grasp how they associate, both at the systemic and atomic level. The network of physical interactions between the proteins of a system, cell or organism, is defined as the interactome. The purpose of this review is to deepen the description of interactomes at different levels of detail: from the molecular structure of complexes to the global topology of the network of interactions. The approaches and techniques applied experimentally and computationally to attain each level are depicted. The limits of each technique and its integration into a model network, the challenges and actual problems of completeness of an interactome, and the reliability of the interactions are reviewed and summarized. Finally, the

application of the current knowledge of protein-protein interactions on modern network medicine and protein function annotation is also explored.

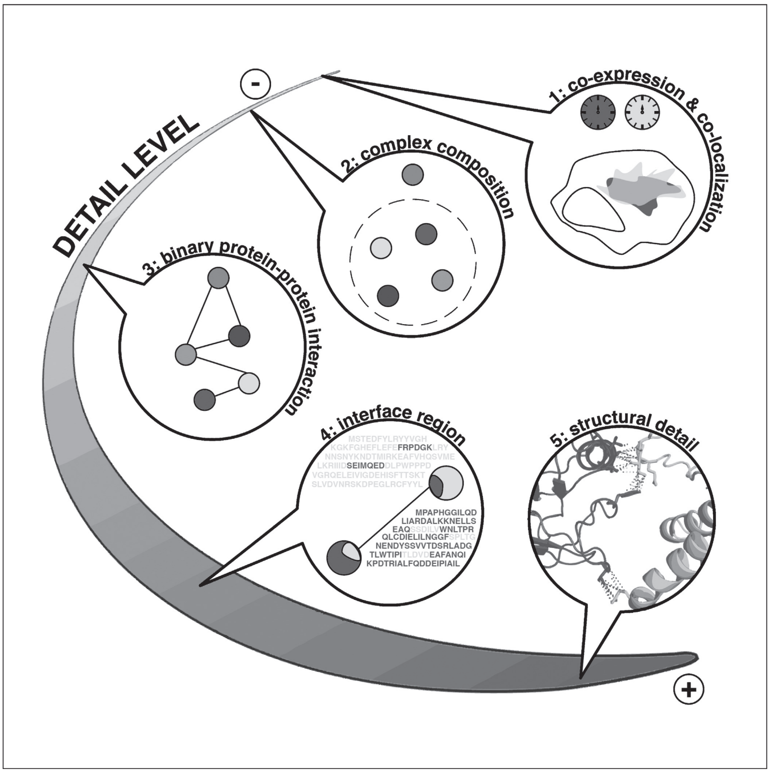
Garcia-Garcia, J., Bonet, J., Guney, E., **Fornes, O.**, Planas-Iglesias, J., & Oliva, B. (2012). **Networks of Protein-Protein Interactions: From Uncertainty to Molecular Details.** *Molecular Informatics*, 31(5), 342-362.

Review

DOI: 10.1002/minf.201200005

Networks of Protein–Protein Interactions: From Uncertainty to Molecular Details

Javier Garcia-Garcia,[a] Jaume Bonet,[a] Emre Guney,[a] Oriol Fornes,[a] Joan Planas,[a] and Baldo Oliva *[a]



7. References

1. Avery OT, MacLeod CM, McCarty M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. J Exp Med. 1944 Feb 1;79(2):137–58.
2. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953 Apr 25;171(4356):737–8.
3. Wilkins MHF, Stokes AR, Wilson HR. Molecular structure of deoxypentose nucleic acids. Nature. 1953 Apr 25;171(4356):738–40.
4. Franklin RE, Gosling RG. Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. Nature. 1953 Jul 25;172(4369):156–7.
5. Watson JD, Crick FH. Genetical implications of the structure of deoxyribonucleic acid. Nature. 1953 May 30;171(4361):964–7.
6. Franklin RE, Gosling RG. Molecular configuration in sodium thymonucleate. Nature. 1953 Apr 25;171(4356):740–1.
7. Gruber TM, Gross CA. Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space. Annu Rev Microbiol. 2003;57(1):441–66.
8. Chakraborty A, Mukherjee S, Chattopadhyay R, Roy S, Chakrabarti S. Conformational Adaptation in the E. coli Sigma 32 Protein in Response to Heat Shock. J Phys Chem B. 2014 May 8;118(18):4793–802.
9. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nat Rev Genet. 2013 Apr;14(4):288–95.
10. Ronshaugen M, Levine M. Visualization of trans-Homolog Enhancer-Promoter Interactions at the Abd-B Hox Locus in the Drosophila Embryo. Dev Cell. 2004 Jan 12;7(6):925–32.
11. Borok MJ, Tran DA, Ho MCW, Drewell RA. Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila. Development. 2010 Jan 1;137(1):5–13.
12. Norris DP, Robertson EJ. Asymmetric and node-specific nodal expression patterns are controlled by two distinct cis-acting regulatory elements. Genes Dev. 1999 Jun 15;13(12):1575–88.
13. Norris DP, Brennan J, Bikoff EK, Robertson EJ. The Foxh1-dependent autoregulatory enhancer controls the level of Nodal signals in the mouse embryo. Development. 2002 Jul 15;129(14):3455–68.
14. Rojas A, Schachterle W, Xu S-M, Martín F, Black BL. Direct transcriptional regulation of Gata4 during early endoderm specification is controlled by FoxA2 binding to an intronic enhancer. Dev Biol. 2010 Oct 15;346(2):346–55.
15. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012 Sep 6;489(7414):75–82.
16. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell. 2013 Apr 11;153(2):307–19.
17. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, et al. Super-Enhancers in the Control of Cell Identity and Disease. Cell. 2013 Jul 11;155(4):934–47.
18. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol. 1961 Jun;3(3):318–56.
19. Ogbourne S, Antalis TM. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. Biochem J. 1998 Apr 1;331(Pt 1):1–14.
20. Sawada S, Scarborough JD, Killeen N, Littman DR. A lineage-specific transcriptional silencer regulates CD4 gene expression during T lymphocyte development. Cell. 1994 Jun 17;77(6):917–29.
21. Dillon N. Gene regulation and large-scale chromatin organization in the nucleus. Chromosome Res. 2006 Feb 1;14(1):117–26.
22. Kadonaga JT. Regulation of RNA Polymerase II Transcription by Sequence-Specific DNA Binding Factors. Cell. 2004 Jan 23;116(2):247–57.
23. Li G, Reinberg D. Chromatin higher-order structures and gene regulation. Curr Opin Genet Dev. 2011 Apr;21(2):175–86.
24. Strahl BD, Allis CD. The language of covalent histone modifications. Nature. 2000 Jan 6;403(6765):41–5.
25. McCulloch SD, Kunkel TA. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. Cell Res. 2008 Jan;18(1):148–61.
26. Stenlund A. Initiation of DNA replication: lessons from viral initiator proteins. Nat Rev Mol Cell Biol. 2003 Oct;4(10):777–85.

27. Bell SP. The origin recognition complex: from simple origins to complex functions. Genes Dev. 2002 Mar 15;16(6):659–72.
28. Takisawa H, Mimura S, Kubota Y. Eukaryotic DNA replication: from pre-replication complex to initiation complex. Curr Opin Cell Biol. 2000 Dec 1;12(6):690–6.
29. Sakabe K, Okazaki R. A unique property of the replicating region of chromosomal DNA. Biochim Biophys Acta BBA - Nucleic Acids Protein Synth. 1966 Dec 21;129(3):651–4.
30. Hayflick L, Moorhead PS. The serial cultivation of human diploid cell strains. Exp Cell Res. 1961 Dec;25(3):585–621.
31. Nandakumar J, Cech TR. Finding the end: recruitment of telomerase to telomeres. Nat Rev Mol Cell Biol. 2013 Feb;14(2):69–82.
32. Richmond TJ, Finch JT, Rushton B, Rhodes D, Klug A. Structure of the nucleosome core particle at 7 |[angst]| resolution. Nature. 1984 Oct 11;311(5986):532–7.
33. Thoma F, Koller T. Influence of histone H1 on chromatin structure. Cell. 1977 Jan 9;12(1):101–7.
34. Woodcock CL. A milestone in the odyssey of higher-order chromatin structure. Nat Struct Mol Biol. 2005 Aug;12(8):639–40.
35. Borland L, Harauz G, Bahr G, van Heel M. Packing of the 30 nm chromatin fiber in the human metaphase chromosome. Chromosoma. 1988;97(2):159–63.
36. Fischle W, Wang Y, Allis CD. Histone and chromatin cross-talk. Curr Opin Cell Biol. 2003 Apr;15(2):172–83.
37. Narlikar GJ, Sundaramoorthy R, Owen-Hughes T. Mechanisms and Functions of ATP-Dependent Chromatin-Remodeling Enzymes. Cell. 2013 Jan 8;154(3):490–503.
38. Schalch T, Duda S, Sargent DF, Richmond TJ. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. Nature. 2005 Jul 7;436(7047):138–41.
39. Peterson CL, Côté J. Cellular machineries for chromosomal DNA repair. Genes Dev. 2004 Mar 15;18(6):602–16.
40. Sinha RP, Häder DP. UV-induced DNA damage and repair: a review. Photochem Photobiol Sci Off J Eur Photochem Assoc Eur Soc Photobiol. 2002 Apr;1(4):225–36.
41. Sedgwick B. Repairing DNA-methylation damage. Nat Rev Mol Cell Biol. 2004 Feb;5(2):148–57.
42. David SS, O'Shea VL, Kundu S. Base-excision repair of oxidative DNA damage. Nature. 2007 Jun 21;447(7147):941–50.
43. Van Gent DC, van der Burg M. Non-homologous end-joining, a sticky affair. Oncogene. 2007;26(56):7731–40.
44. Li G-M. Mechanisms and functions of DNA mismatch repair. Cell Res. 2008 Jan;18(1):85–98.
45. Li X, Heyer W-D. Homologous recombination in DNA repair and DNA damage tolerance. Cell Res. 2008 Jan;18(1):99–113.
46. Weiner A, Zauberman N, Minsky A. Recombinational DNA repair in a cellular context: a search for the homology search. Nat Rev Microbiol. 2009 Oct;7(10):748–55.
47. Polo SE, Jackson SP. Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. Genes Dev. 2011 Mar 1;25(5):409–33.
48. Muller HJ. The relation of recombination to mutational advance. Mutat Res Mol Mech Mutagen. 1964 May;1(1):2–9.
49. Xu Z, Zan H, Pone EJ, Mai T, Casali P. Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. Nat Rev Immunol. 2012 Jul;12(7):517–31.
50. Furney SJ, Higgins DG, Ouzounis CA, López-Bigas N. Structural and functional properties of genes involved in human cancer. BMC Genomics. 2006 Jan 11;7(1):3.
51. Boyadjiev S, Jabs E. Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. Clin Genet. 2000 Apr 1;57(4):253–66.
52. Van Nimwegen E. Scaling laws in the functional content of genomes. Trends Genet. 2003 Sep;19(9):479–84.
53. Gray PA, Fu H, Luo P, Zhao Q, Yu J, Ferrari A, et al. Mouse Brain Organization Revealed Through Direct Genome-Scale TF Expression Analysis. Science. 2004 Dec 24;306(5705):2255–7.
54. Messina DN, Glasscock J, Gish W, Lovett M. An ORFeome-based Analysis of Human Transcription Factor Genes and the Construction of a Microarray to Interrogate Their Expression. Genome Res. 2004 Oct 15;14(10b):2041–7.
55. Reece-Hoyes JS, Deplancke B, Shingles J, Grove CA, Hope IA, Walhout AJ. A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. Genome Biol. 2005 Dec 30;6(13):R110.

56. Adryan B, Teichmann SA. FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster. Bioinformatics. 2006 Jun 15;22(12):1532–3.
57. Ho S-W, Jona G, Chen CTL, Johnston M, Snyder M. Linking DNA-binding proteins to their recognition sequences by using protein microarrays. Proc Natl Acad Sci. 2006 Jun 27;103(26):9940–5.
58. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009 Apr;10(4):252–63.
59. Kuras L, Struhl K. Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme. Nature. 1999 Jun 10;399(6736):609–13.
60. Cang Y, Auble DT, Prelich G. A new regulatory domain on the TATA-binding protein. EMBO J. 1999 Dec 1;18(23):6662–71.
61. Kraemer SM, Ranallo RT, Ogg RC, Stargell LA. TFIIA Interacts with TFIID via Association with TATA-Binding Protein and TAF40. Mol Cell Biol. 2001 Mar 1;21(5):1737–46.
62. Deng W, Roberts SGE. TFIIB and the regulation of transcription by RNA polymerase II. Chromosoma. 2007 Oct 1;116(5):417–29.
63. Bushnell DA, Westover KD, Davis RE, Kornberg RD. Structural Basis of Transcription: An RNA Polymerase II-TFIIB Cocrystal at 4.5 Angstroms. Science. 2004 Feb 13;303(5660):983–8.
64. Kostrewa D, Zeller ME, Armache K-J, Seizl M, Leike K, Thomm M, et al. RNA polymerase II–TFIIB structure and mechanism of transcription initiation. Nature. 2009 Nov 19;462(7271):323–30.
65. Liu X, Bushnell DA, Wang D, Calero G, Kornberg RD. Structure of an RNA Polymerase II–TFIIB Complex and the Transcription Initiation Mechanism. Science. 2010 Jan 8;327(5962):206–9.
66. Compe E, Egly J-M. TFIIH: when transcription met DNA repair. Nat Rev Mol Cell Biol. 2012 Jun;13(6):343–54.
67. Kim Y, Geiger JH, Hahn S, Sigler PB. Crystal structure of a yeast TBP/TATA-box complex. Nature. 1993 Oct 7;365(6446):512–20.
68. Nikolov DB, Chen H, Halay ED, Usheva AA, Hisatake K, Lee DK, et al. Crystal structure of a TFIIB–TBP–TATA-element ternary complex. Nature. 1995 Sep 14;377(6545):119–28.
69. Tan S, Hunziker Y, Sargent DF, Richmond TJ. Crystal structure of a yeast TFIIA/TBP/DNA complex. Nature. 1996 May 9;381(6578):127–34.
70. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—A visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605–12.
71. Baumann M, Pontiller J, Ernst W. Structure and Basal Transcription Complex of RNA Polymerase II Core Promoters in the Mammalian Genome: An Overview. Mol Biotechnol. 2010 Jul 1;45(3):241–7.
72. Basehoar AD, Zanton SJ, Pugh BF. Identification and Distinct Regulation of Yeast TATA Box-Containing Genes. Cell. 2004 May 3;116(5):699–709.
73. Patikoglou GA, Kim JL, Sun L, Yang S-H, Kodadek T, Burley SK. TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. Genes Dev. 1999 Dec 15;13(24):3217–30.
74. Piskacek M. Common Transactivation Motif 9aaTAD recruits multiple general co-activators TAF9, MED15, CBP and p300. Nat Preced. 2009 Nov 5; Available from: http://precedings.nature.com/documents/3488/version/2
75. Rutherford JC, Ojeda L, Balk J, Mühlenhoff U, Lill R, Winge DR. Activation of the Iron Regulon by the Yeast Aft1/Aft2 Transcription Factors Depends on Mitochondrial but Not Cytosolic Iron-Sulfur Protein Biogenesis. J Biol Chem. 2005 Mar 18;280(11):10135–40.
76. Hilger-Eversheim K, Moser M, Schorle H, Buettner R. Regulatory roles of AP-2 transcription factors in vertebrate development, apoptosis and cell-cycle control. Gene. 2000 Dec 30;260(1–2):1–12.
77. Wilsker D, Patsialou A, Dallas PB, Moran E. ARID Proteins: A Diverse Family of DNA Binding Proteins Implicated in the Control of Cell Growth, Differentiation, and Development. Cell Growth Differ. 2002 Mar 1;13(3):95.
78. Jones S. An overview of the basic helix-loop-helix proteins. Genome Biol. 2004 May 28;5(6):226.
79. Blank V, Andrews NC. The Maf transcription factors: regulators of differentiation. Trends Biochem Sci. 1997 Nov;22(11):437–41.

80. Don J, Stelzer G. The expanding family of CREB/CREM transcription factors that are involved with spermatogenesis. Mol Cell Endocrinol. 2002 Feb 22;187(1–2):115–24.
81. Hai T, Wolfgang CD, Marsee DK, Allen AE, Sivaprasad U. ATF3 and stress responses. Gene Expr. 1999;7(4-6):321–35.
82. Regulation of Steroidogenesis and the Steroidogenic Acute Regulatory Protein by a Member of the cAMP Response-Element Binding Protein Family. Mol Endocrinol. 2002 Jan 1;16(1):184–99.
83. Wolfe SA, Nekludova L, Pabo CO. DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct. 2000;29(1):183–212.
84. Müller H, Helin K. The E2F transcription factors: key regulators of cell proliferation. Biochim Biophys Acta. 2000 Feb 14;1470(1):M1–12.
85. Sharrocks AD. The ETS-domain transcription factor family. Nat Rev Mol Cell Biol. 2001 Nov;2(11):827–37.
86. Lehmann OJ, Sowden JC, Carlsson P, Jordan T, Bhattacharya SS. Fox's in development and disease. Trends Genet. 2003 Jun;19(6):339–44.
87. Bresnick EH, Katsumura KR, Lee H-Y, Johnson KD, Perkins AS. Master regulatory GATA transcription factors: mechanistic principles and emerging links to hematologic malignancies. Nucleic Acids Res. 2012 Apr 5;gks281.
88. Bürglin TR. Homeodomain subtypes and functional diversity. In: Hughes TR, editor. A Handbook of Transcription Factors [Internet]. Springer Netherlands; 2011 [cited 2014 Jul 9]. p. 95–122. Available from: http://link.springer.com/chapter/10.1007/978-90-481-9069-0_5
89. Tamura T, Yanai H, Savitsky D, Taniguchi T. The IRF Family Transcription Factors in Immunity and Oncogenesis. Annu Rev Immunol. 2008;26(1):535–84.
90. Shore P, Sharrocks AD. The MADS-box family of transcription factors. Eur J Biochem FEBS. 1995 Apr 1;229(1):1–13.
91. Myb binding proteins: regulators and cohorts in transformation. 1999. Available: http://www.nature.com/onc/journal/v18/n19/full/1202726a.html
92. Olsen AN, Ernst HA, Leggio LL, Skriver K. NAC transcription factors: structurally distinct, functionally diverse. Trends Plant Sci. 2005 Jan 2;10(2):79–87.
93. Robinson-Rechavi M, Garcia HE, Laudet V. The nuclear receptor superfamily. J Cell Sci. 2003 Feb 15;116(4):585–6.
94. Dahl E, Koseki H, Balling R. Pax genes and organogenesis. BioEssays. 1997 Sep 1;19(9):755–65.
95. Latchman DS. POU family transcription factors in the nervous system. J Cell Physiol. 1999 May;179(2):126–33.
96. Andersen B, Rosenfeld MG. POU domain factors in the neuroendocrine system: lessons from developmental biology provide insights into human disease. Endocr Rev. 2001 Feb;22(1):2–35.
97. Massagué J, Seoane J, Wotton D. Smad transcription factors. Genes Dev. 2005 Dec 1;19(23):2783–810.
98. Wegner M. From head to toes: The multiple facets of Sox proteins. Nucleic Acids Res. 1999 Mar 1;27(6):1409–20.
99. Wilson V, Conlon FL. The T-box family. Genome Biol. 2002;3(6):REVIEWS3008.
100. De Nadal E, Ammerer G, Posas F. Controlling gene expression in response to stress. Nat Rev Genet. 2011 Dec;12(12):833–45.
101. Pandey SP, Somssich IE. The Role of WRKY Transcription Factors in Plant Immunity. Plant Physiol. 2009 Aug 1;150(4):1648–55.
102. Laity JH, Lee BM, Wright PE. Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol. 2001 Feb 1;11(1):39–46.
103. Poor CB, Wegner SV, Li H, Dlouhy AC, Schuermann JP, Sanishvili R, et al. Molecular mechanism and structure of the Saccharomyces cerevisiae iron regulator Aft2. Proc Natl Acad Sci. 2014 Mar 18;111(11):4043–8.
104. Allen MD, Yamasaki K, Ohme□Takagi M, Tateno M, Suzuki M. A novel mode of DNA recognition by a β□sheet revealed by the solution structure of the GCC□box binding domain in complex with DNA. EMBO J. 1998 Sep 15;17(18):5484–96.
105. Iwahara J, Iwahara M, Daughdrill GW, Ford J, Clubb RT. The structure of the Dead ringer–DNA complex reveals how AT□rich interaction domains (ARIDs) recognize DNA. EMBO J. 2002 Mar 1;21(5):1197–209.
106. Nair SK, Burley SK. X-Ray Structures of Myc-Max and Mad-Max Recognizing DNA. Cell. 2003 Jan 24;112(2):193–205.
107. Glover JNM, Harrison SC. Crystal structure of the heterodimeric bZIP transcription factor c-Fos–c-Jun bound to DNA. Nature. 1995 Jan 19;373(6511):257–61.

108. Pavletich NP, Pabo CO. Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. Science. 1993 Sep 24;261(5129):1701–7.
109. Zheng N, Fraenkel E, Pabo CO, Pavletich NP. Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F–DP. Genes Dev. 1999 Mar 15;13(6):666–74.
110. Kodandapani R, Pio F, Ni C-Z, Piccialli G, Klemsz M, McKercher S, et al. A new pattern for helix–turn–helix recognition revealed by the PU.l ETS–domain–DNA complex. Nature. 1996 Apr 4;380(6573):456–60.
111. Clark KL, Halay ED, Lai E, Burley SK. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. Nature. 1993 Jul 29;364(6436):412–20.
112. Omichinski JG, Clore GM, Schaad O, Felsenfeld G, Trainor C, Appella E, et al. NMR structure of a specific DNA complex of Zn-containing DNA binding domain of GATA-1. Science. 1993 Jul 23;261(5120):438–46.
113. Kissinger CR, Liu BS, Martin-Blanco E, Kornberg TB, Pabo CO. Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: a framework for understanding homeodomain-DNA interactions. Cell. 1990 Nov 2;63(3):579–90.
114. Fujii Y, Shimizu T, Kusumoto M, Kyogoku Y, Taniguchi T, Hakoshima T. Crystal structure of an IRF□DNA complex reveals novel DNA recognition and cooperative binding to a tandem repeat of core sequences. EMBO J. 1999 Sep 15;18(18):5028–41.
115. Tan S, Richmond TJ. Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. Nature. 1998 Feb 12;391(6668):660–6.
116. Konig P, Giraldo R, Chapman L, Rhodes D. The crystal structure of the DNA-binding domain of yeast RAP1 in complex with telomeric DNA. Cell. 1996 Apr 5;85(1):125–36.
117. Welner DH, Lindemose S, Grossmann JG, Møllegaard NE, Olsen AN, Helgstrand C, et al. DNA binding by the plant-specific NAC transcription factors in crystal and solution: a firm link to WRKY and GCM transcription factors. Biochem J. 2012 Jun 15;444(3):395–404.
118. Chandra V, Huang P, Hamuro Y, Raghuram S, Wang Y, Burris TP, et al. Structure of the intact PPAR-γ–RXR-α nuclear receptor complex on DNA. Nature. 2008 Oct 29;456(7220):350–6.
119. Xu HE, Rould MA, Xu W, Epstein JA, Maas RL, Pabo CO. Crystal structure of the human Pax6 paired domain-DNA complex reveals specific roles for the linker region and carboxy-terminal subdomain in DNA binding. Genes Dev. 1999 May 15;13(10):1263–75.
120. Klemm JD, Rould MA, Aurora R, Herr W, Pabo CO. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. Cell. 1994 Apr 8;77(1):21–32.
121. Shi Y, Wang YF, Jayaraman L, Yang H, Massagué J, Pavletich NP. Crystal structure of a Smad MH1 domain bound to DNA: insights on DNA binding in TGF-beta signaling. Cell. 1998 Sep 4;94(5):585–94.
122. Werner MH, Huth JR, Gronenborn AM, Clore GM. Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. Cell. 1995 Jun 2;81(5):705–14.
123. Müller CW, Herrmann BG. Crystallographic structure of the T domain–DNA complex of the Brachyury transcription factor. Nature. 1997 Oct 23;389(6653):884–8.
124. Yamasaki K, Kigawa T, Watanabe S, Inoue M, Yamasaki T, Seki M, et al. Structural Basis for Sequence-specific DNA Recognition by an Arabidopsis WRKY Transcription Factor. J Biol Chem. 2012 Mar 2;287(10):7683–91.
125. Pavletich NP, Pabo CO. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science. 1991 May 10;252(5007):809–17.
126. Wang GL, Jiang BH, Rue EA, Semenza GL. Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension. Proc Natl Acad Sci. 1995 Jun 6;92(12):5510–4.
127. Mitchell TJ, John S. Signal transducer and activator of transcription (STAT) signalling and T-cell lymphomas. Immunology. 2005 Mar;114(3):301–12.
128. MacDonald BT, Tamai K, He X. Wnt/β-Catenin Signaling: Components, Mechanisms, and Diseases. Dev Cell. 2009 Jul 21;17(1):9–26.
129. Fujii Y, Shimizu T, Toda T, Yanagida M, Hakoshima T. Structural basis for the diversity of DNA recognition by bZIP transcription factors. Nat Struct Mol Biol. 2000 Oct;7(10):889–93.
130. Helsen C, Kerkhofs S, Clinckemalie L, Spans L, Laurent M, Boonen S, et al. Structural basis for nuclear hormone receptor DNA binding. Mol Cell Endocrinol. 2012 Jan 30;348(2):411–7.
131. Marianayagam NJ, Sunde M, Matthews JM. The power of two: protein dimerization in biology. Trends Biochem Sci. 2004 Nov;29(11):618–25.

132. Chikka MR, McCabe DD, Tyra HM, Rutkowski DT. C/EBP homologous protein (CHOP) contributes to suppression of metabolic genes during endoplasmic reticulum stress in the liver. J Biol Chem. 2013 Feb 8;288(6):4405–15.
133. Panne D. The enhanceosome. Curr Opin Struct Biol. 2008 Apr;18(2):236–42.
134. Michnick SW. Exploring protein interactions by interaction-induced folding of proteins from complementary peptide fragments. Curr Opin Struct Biol. 2001 Aug 1;11(4):472–7.
135. Bruckner A, Polge C, Lentze N, Auerbach D, Schlattner U. Yeast two-hybrid, a powerful tool for systems biology. Int J Mol Sci. 2009 Jun 18;10(6):2763–88.
136. Xu X, Song Y, Li Y, Chang J, zhang H, An L. The tandem affinity purification method: an efficient system for protein complex purification and protein interaction identification. Protein Expr Purif. 2010 Aug;72(2):149–56.
137. Hu C-D, Chinenov Y, Kerppola TK. Visualization of Interactions among bZIP and Rel Family Proteins in Living Cells Using Bimolecular Fluorescence Complementation. Mol Cell. 2002 Apr;9(4):789–98.
138. Magliery TJ, Wilson CGM, Pan W, Mishler D, Ghosh I, Hamilton AD, et al. Detecting Protein−Protein Interactions with a Green Fluorescent Protein Fragment Reassembly Trap: Scope and Mechanism. J Am Chem Soc. 2005 Jan 1;127(1):146–57.
139. Day RN, Periasamy A, Schaufele F. Fluorescence resonance energy transfer microscopy of localized protein interactions in the living cell nucleus. Methods San Diego Calif. 2001 Sep;25(1):4–18.
140. Xu Y, Piston DW, Johnson CH. A bioluminescence resonance energy transfer (BRET) system: Application to interacting circadian clock proteins. Proc Natl Acad Sci. 1999 Jan 5;96(1):151–6.
141. MacBeath G, Schreiber SL. Printing proteins as microarrays for high-throughput function fetermination. Science. 2000 Sep 8;289(5485):1760–3.
142. Boozer C, Kim G, Cong S, Guan H, Londergan T. Looking towards label-free biomolecular interaction analysis in a high-throughput format: a review of new surface plasmon resonance technologies. Curr Opin Biotechnol. 2006 Aug;17(4):400–5.
143. Lemmens I, Eyckerman S, Zabeau L, Catteeuw D, Vertenten E, Verschueren K, et al. Heteromeric MAPPIT: a novel strategy to study modification□dependent protein–protein interactions in mammalian cells. Nucleic Acids Res. 2003 Jul 15;31(14):e75–e75.
144. Russ WP, Engelman DM. TOXCAT: a measure of transmembrane helix association in a biological membrane. Proc Natl Acad Sci. 1999 Feb 2;96(3):863–8.
145. Söderberg O, Gullberg M, Jarvius M, Ridderstråle K, Leuchowius K-J, Jarvius J, et al. Direct observation of individual endogenous protein complexes in situ by proximity ligation. Nat Methods. 2006 Dec;3(12):995–1000.
146. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 2004 Jan 1;32(suppl 1):D449–51.
147. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, et al. The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res. 2005 Jan 1;33(suppl 1):D418–24.
148. Chatr-aryamontri A, Breitkreutz B-J, Heinicke S, Boucher L, Winter A, Stark C, et al. The BioGRID interaction database: 2013 update. Nucleic Acids Res. 2013 Jan 1;41(D1):D816–23.
149. Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database—2009 update. Nucleic Acids Res. 2009 Jan 1;37(suppl 1):D767–72.
150. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014 Jan 1;42(D1):D358–63.
151. Garcia-Garcia J, Bonet J, Guney E, Fornes O, Planas J, Oliva B. Networks of protein-protein interactions: from uncertainty to molecular details. Mol Inform. 2012;31(5):342–62.
152. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. 1988 Jan 29;239(4839):487–91.
153. Schmitz A, Galas DJ. The interaction of RNA polymerase and lac repressor with the lac control region. Nucleic Acids Res. 1979 Jan;6(1):111–37.
154. Tullius TD. DNA footprinting with hydroxyl radical. Nature. 1988 Apr 14;332(6165):663–4.
155. Becker MM, Lesser D, Kurpiewski M, Baranger A, Jen-Jacobson L. “Ultraviolet footprinting” accurately maps sequence-specific contacts and DNA kinking in the EcoRI endonuclease-DNA complex. Proc Natl Acad Sci U S A. 1988 Sep;85(17):6247–51.

156. Bowen B, Steinberg J, Laemmli UK, Weintraub H. The detection of DNA-binding proteins by protein blotting. Nucleic Acids Res. 1980 Jan 11;8(1):1–20.
157. Hellman LM, Fried MG. Electrophoretic Mobility Shift Assay (EMSA) for Detecting Protein-Nucleic Acid Interactions. Nat Protoc. 2007;2(8):1849–61.
158. Collas P. The Current State of Chromatin Immunoprecipitation. Mol Biotechnol. 2010 May 1;45(1):87–100.
159. Vogel MJ, Peric-Hupkes D, van Steensel B. Detection of in vivo protein–DNA interactions using DamID in mammalian cells. Nat Protoc. 2007 Jun;2(6):1467–78.
160. Buck MJ, Lieb JD. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics. 2004 Mar;83(3):349–60.
161. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009 Oct;10(10):669–80.
162. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science. 1990 Aug 3;249(4968):505–10.
163. Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief Bioinform. 2008 Jul 1;9(4):326–32.
164. Roulet E, Busso S, Camargo AA, Simpson AJG, Mermod N, Bucher P. High-throughput SELEX–SAGE method for quantitative modeling of transcription-factor binding sites. Nat Biotechnol. 2002 Aug;20(8):831–5.
165. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 2006 Nov;24(11):1429–35.
166. Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc. 2009 Mar;4(3):393–411.
167. Reece-Hoyes JS, Diallo A, Lajoie B, Kent A, Shrestha S, Kadreppa S, et al. Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. Nat Methods. 2011 Dec;8(12):1059–64.
168. Hu S, Xie Z, Blackshaw S, Qian J, Zhu H. Characterization of Protein–DNA Interactions Using Protein Microarrays. Cold Spring Harb Protoc. 2011 May 1;2011(5):pdb.prot5614.
169. Hu S, Xie Z, Onishi A, Yu X, Jiang L, Lin J, et al. Profiling the Human Protein-DNA Interactome Reveals ERK2 as a Transcriptional Repressor of Interferon Signaling. Cell. 2009 Oct 30;139(3):610–22.
170. Tolhuis B, Palstra R-J, Splinter E, Grosveld F, de Laat W. Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. Mol Cell. 2002 Jan 12;10(6):1453–65.
171. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing Chromosome Conformation. Science. 2002 Feb 15;295(5558):1306–11.
172. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nat Genet. 2006 Nov;38(11):1348–54.
173. Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet. 2006 Nov;38(11):1341–7.
174. Dostie J, Dekker J. Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc. 2007 Apr;2(4):988–1002.
175. Keene MA, Elgin SCR. Micrococcal nuclease as a probe of DNA sequence organization and chromatin structure. Cell. 1981 Nov;27(1, Part 2):57–64.
176. Staynov DZ. DNase I footprinting of the nucleosome in whole nuclei. Biochem Biophys Res Commun. 2008 Jul 18;372(1):226–9.
177. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, et al. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Methods. 2006 Jul;3(7):511–8.
178. Vierstra J, Wang H, John S, Sandstrom R, Stamatoyannopoulos JA. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. Nat Methods. 2014 Jan;11(1):66–72.
179. Kasinathan S, Orsi GA, Zentner GE, Ahmad K, Henikoff S. High-resolution mapping of transcription factor binding sites on native chromatin. Nat Methods. 2014 Feb;11(2):203–9.
180. Murtha M, Tokcaer-Keskin Z, Tang Z, Strino F, Chen X, Wang Y, et al. FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. Nat Methods. 2014 May;11(5):559–65.

181. Bowman GD, O'Donnell M, Kuriyan J. Structural analysis of a eukaryotic sliding DNA clamp–clamp loader complex. Nature. 2004 Jun 17;429(6993):724–30.

182. Kelch BA, Makino DL, O'Donnell M, Kuriyan J. How a DNA Polymerase Clamp Loader Opens a Sliding Clamp. Science. 2011 Dec 23;334(6063):1675–80.

183. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–42.

184. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein-protein interfaces. J Mol Biol. 2004 Feb 27;336(4):943–55.

185. Kalodimos CG, Biris N, Bonvin AMJJ, Levandoski MM, Guennuegues M, Boelens R, et al. Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. Science. 2004 Jul 16;305(5682):386–9.

186. Yu H. Extending the size limit of protein nuclear magnetic resonance. Proc Natl Acad Sci. 1999 Jan 19;96(2):332–4.

187. Putnam CD, Hammel M, Hura GL, Tainer JA. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q Rev Biophys. 2007;40(03):191–285.

188. Günther S, Rother K, Frömmel C. Molecular flexibility in protein–DNA interactions. Biosystems. 2006 Aug;85(2):126–36.

189. Russel D, Lasker K, Webb B, Velázquez-Muriel J, Tjioe E, Schneidman-Duhovny D, et al. Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. PLoS Biol. 2012 Jan 17;10(1):e1001244.

190. Planas-Iglesias J, Bonet J, A. M, Feliu E, Gursoy A, Oliva B. Structural Bioinformatics of Proteins: Predicting the Tertiary and Quaternary Structure of Proteins from Sequence. In: Cai W, editor. Protein-Protein Interactions - Computational and Experimental Tools [Internet]. InTech; 2012 [cited 2013 Jul 22]. Available from: http://www.intechopen.com/books/protein-protein-interactions-computational-and-experimental-tools/structural-bioinformatics-of-proteins-predicting-the-tertiary-and-quaternary-structure-of-proteins-f

191. Eswar N, Webb B, Marti-Renom MA, Madhusudhan M s., Eramian D, Shen M, et al. Comparative Protein Structure Modeling Using Modeller. Current Protocols in Bioinformatics [Internet]. John Wiley & Sons, Inc.; 2006 [cited 2013 Sep 24]. Available from: http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi0506s15/abstract

192. Lu X-J, Olson WK. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc. 2008 Jul;3(7):1213–27.

193. Chen C-Y, Chien T-Y, Lin C-K, Lin C-W, Weng Y-Z, Chang DT-H. Predicting Target DNA Sequences of DNA-Binding Proteins Based on Unbound Structures. PLoS ONE. 2012 Feb 1;7(2):e30446.

194. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005 Jan 1;33(7):2302–9.

195. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999 Feb 1;12(2):85–94.

196. Knegtel RMA, Antoon J, Rullmann C, Boelens R, Kaptein R. MONTY: a Monte Carlo approach to protein-DNA recognition. J Mol Biol. 1994 Jan 7;235(1):318–24.

197. Poulain P, Saladin A, Hartmann B, Prévost C. Insights on protein-DNA recognition by coarse grain modelling. J Comput Chem. 2008;29(15):2582–92.

198. Dijk M van, Bonvin AMJJ. Pushing the limits of what is achievable in protein–DNA docking: benchmarking HADDOCK's performance. Nucleic Acids Res. 2010 Sep 1;38(17):5634–47.

199. Parisien M, Freed KF, Sosnick TR. On Docking, Scoring and Assessing Protein-DNA Complexes in a Rigid-Body Framework. PLoS ONE. 2012 Feb 29;7(2):e32647.

200. Dijk M van, Visscher KM, Kastritis PL, Bonvin AMJJ. Solvated protein–DNA docking using HADDOCK. J Biomol NMR. 2013 May 1;56(1):51–63.

201. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell. 2014 Sep 11;158(6):1431–43.

202. Sander JD, Maeder ML, Reyon D, Voytas DF, Joung JK, Dobbs D. ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W462–8.

203. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics. 2004 Mar 1;20(4):477–86.

204. Yu X, Cao J, Cai Y, Shi T, Li Y. Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. J Theor Biol. 2006 May 21;240(2):175–84.

205. Kumar M, Gromiha MM, Raghava GP. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics. 2007 Nov 27;8(1):463.

206. Nimrod G, Schushan M, Szilágyi A, Leslie C, Ben-Tal N. iDBPs: a web server for the identification of DNA binding proteins. Bioinformatics. 2010 Mar 1;26(5):692–3.

207. Shazman S, Celniker G, Haber O, Glaser F, Mandel-Gutfreund Y. Patch Finder Plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W526–30.

208. Gao M, Skolnick J. DBD-Hunter: a knowledge-based method for the prediction of DNA–protein interactions. Nucleic Acids Res. 2008 Jul 1;36(12):3978–92.

209. Gao M, Skolnick J. A threading-based method for the prediction of DNA-binding proteins with application to the human genome. PLoS Comput Biol. 2009 Nov;5(11):e1000567.

210. Zhao H, Yang Y, Zhou Y. Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. Bioinformatics. 2010 Aug 1;26(15):1857–63.

211. Luscombe NM, Thornton JM. Protein–DNA Interactions: Amino Acid Conservation and the Effects of Mutations on Binding Specificity. J Mol Biol. 2002 Jul 26;320(5):991–1009.

212. Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res. 2006 Jul 1;34(suppl 2):W243–8.

213. Hwang S, Gou Z, Kuznetsov IB. DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics. 2007 Mar 1;23(5):634–6.

214. Carson MB, Langlois R, Lu H. NAPS: a residue-level nucleic acid-binding prediction server. Nucleic Acids Res. 2010 Jul 1;38(suppl 2):W431–5.

215. Wang L, Huang C, Yang MQ, Yang JY. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol. 2010 May 28;4(Suppl 1):S3.

216. Si J, Zhang Z, Lin B, Schroeder M, Huang B. MetaDBSite: a meta approach to improve protein DNA-binding sites prediction [Internet]. BioMed Central Ltd; 2011 Jun [cited 2013 Oct 16] p. S7. Report No.: Suppl 1. Available from: http://www.biomedcentral.com/1752-0509/5/S1/S7/abstract

217. Tjong H, Zhou H-X. DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. Nucleic Acids Res. 2007 Mar 1;35(5):1465–77.

218. Ozbek P, Soner S, Erman B, Haliloglu T. DNABINDPROT: fluctuation-based predictor of DNA-binding residues within a network of interacting residues. Nucleic Acids Res. 2010 Jul 1;38(suppl 2):W417–23.

219. Chen YC, Wright JD, Lim C. DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry. Nucleic Acids Res. 2012 May 31;40(W1):W249–56.

220. Das MK, Dai H-K. A survey of DNA motif finding algorithms. BMC Bioinformatics. 2007 Nov 1;8(Suppl 7):S21.

221. Lin W-Z, Fang J-A, Xiao X, Chou K-C. iDNA-Prot: identification of DNA binding proteins using random forest with grey model. PLoS ONE. 2011 Sep 15;6(9):e24756.

222. Chu W-Y, Huang Y-F, Huang C-C, Cheng Y-S, Huang C-K, Oyang Y-J. ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors. Nucleic Acids Res. 2009 Jul 1;37(suppl 2):W396–401.

223. Gabdoulline R, Eckweiler D, Kel A, Stegmaier P. 3DTF: a web server for predicting transcription factor PWMs using 3D structure-based energy calculations. Nucleic Acids Res. 2012 Jul 1;40(W1):W180–5.

224. Chien T-Y, Lin C-K, Lin C-W, Weng Y-Z, Chen C-Y, Chang DT-H. DBD2BS: connecting a DNA-binding protein with its binding sites. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W173–9.

225. Xu B, Yang Y, Liang H, Zhou Y. An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. Proteins Struct Funct Bioinforma. 2009;76(3):718–30.

226. Lin C-K, Chen C-Y. PiDNA: predicting protein–DNA interactions with structural models. Nucleic Acids Res. 2013 Jul 1;41(W1):W523–30.

227. Wang L, Yang MQ, Yang JY. Prediction of DNA-binding residues from protein sequence information using random forests. BMC Genomics. 2009 Jul 7;10(Suppl 1):S1.

228. Wu J, Liu H, Duan X, Ding Y, Wu H, Bai Y, et al. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics. 2009 Jan 1;25(1):30–5.

229. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics. 2005 Feb 19;6(1):33.

230. Li T, Li Q-Z, Liu S, Fan G-L, Zuo Y-C, Peng Y. PreDNA: accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information. Bioinformatics. 2013 Mar 15;29(6):678–85.

231. Christensen RG, Enuameh MS, Noyes MB, Brodsky MH, Wolfe SA, Stormo GD. Recognition models to predict DNA-binding specificities of homeodomain proteins. Bioinformatics. 2012 Jun 15;28(12):i84–9.

232. Gupta A, Christensen RG, Bell HA, Goodwin M, Patel RY, Pandey M, et al. An improved predictive recognition model for Cys2-His2 zinc finger proteins. Nucleic Acids Res. 2014 Feb 12;gku132.

233. Persikov AV, Singh M. De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. Nucleic Acids Res. 2014 Jan;42(1):97–108.

234. Angarica VE, Pérez AG, Vasconcelos AT, Collado-Vides J, Contreras-Moreira B. Prediction of TF target sites based on atomistic models of protein-DNA complexes. BMC Bioinformatics. 2008 Oct 16;9(1):436.

235. Liu Z, Guo J-T, Li T, Xu Y. Structure-based prediction of transcription factor binding sites using a protein-DNA docking approach. Proteins Struct Funct Bioinforma. 2008;72(4):1114–24.

236. Alamanova D, Stegmaier P, Kel A. Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. BMC Bioinformatics. 2010 May 3;11(1):225.

237. AlQuraishi M, McAdams HH. Three enhancements to the inference of statistical protein-DNA potentials. Proteins Struct Funct Bioinforma. 2013;81(3):426–42.

238. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2014 Jan 1;42(D1):D142–7.

239. Yang J-H, Li J-H, Jiang S, Zhou H, Qu L-H. ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. Nucleic Acids Res. 2013 Jan 1;41(D1):D177–87.

240. Robasky K, Bulyk ML. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein–DNA interactions. Nucleic Acids Res. 2011 Jan 1;39(suppl 1):D124–8.

241. Teixeira MC, Monteiro PT, Guerreiro JF, Gonçalves JP, Mira NP, Santos SC dos, et al. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. Nucleic Acids Res. 2014 Jan 1;42(D1):D161–6.

242. Boer CG de, Hughes TR. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. Nucleic Acids Res. 2011 Nov 18;gkr993.

243. Gallo SM, Gerrard DT, Miner D, Simich M, Soye BD, Bergman CM, et al. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. Nucleic Acids Res. 2011 Jan 1;39(suppl 1):D118–23.

244. Shazman S, Lee H, Socol Y, Mann RS, Honig B. OnTheFly: a database of Drosophila melanogaster transcription factors and their binding sites. Nucleic Acids Res. 2014 Jan 1;42(D1):D167–71.

245. Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. hPDI: a database of experimental human protein–DNA interactions. Bioinformatics. 2010 Jan 15;26(2):287–9.

246. Moreland RT, Ryan JF, Pan C, Baxevanis AD. The Homeodomain Resource: a comprehensive collection of sequence, structure, interaction, genomic and functional information on the homeodomain protein family. Database. 2009 Apr 28;2009(0):bap004–bap004.

247. Fu F, Voytas DF. Zinc Finger Database (ZiFDB) v2.0: a comprehensive database of C2H2 zinc fingers and engineered zinc finger arrays. Nucleic Acids Res. 2013 Jan 1;41(D1):D452–5.

248. Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucleic Acids Res. 2006 Jan 1;34(suppl 1):D204–6.

249. Kirsanov DD, Zanegina ON, Aksianov EA, Spirin SA, Karyagina AS, Alexeevski AV. NPIDB: nucleic acid--protein interaction database. Nucleic Acids Res. 2012 Nov 27;41(D1):D517–23.

250. Contreras-Moreira B. 3D-footprint: a database for the structural analysis of protein–DNA complexes. Nucleic Acids Res. 2010 Jan 1;38(suppl 1):D91–7.

251. Lee S, Blundell TL. BIPA: a database for protein–nucleic acid interaction in 3D structures. Bioinformatics. 2009 Jun 15;25(12):1559–60.

252. Turner D, Kim R, Guo J. TFinDit: transcription factor-DNA interaction data depository. BMC Bioinformatics. 2012 Sep 3;13(1):220.

253. Sebastian A, Contreras-Moreira B. footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. Bioinformatics. 2014 Jan 15;30(2):258–65.
254. Hoffman MM, Khrapov MA, Cox JC, Yao J, Tong L, Ellington AD. AANT: the Amino Acid–Nucleotide Interaction Database. Nucleic Acids Res. 2004 Jan 1;32(suppl 1):D174–81.
255. Norambuena T, Melo F. The Protein-DNA Interface database. BMC Bioinformatics. 2010 May 18;11(1):262.
256. Yang L, Zhou T, Dror I, Mathelier A, Wasserman WW, Gordân R, et al. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. 2014 Jan 1;42(D1):D148–55.
257. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, et al. Protein function annotation by homology-based inference. Genome Biol. 2009 Feb 2;10(2):207.
258. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, et al. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. Cell. 2008 Jun 27;133(7):1266–76.
259. Espadaler J, Aragüés R, Eswar N, Marti-Renom MA, Querol E, Avilés FX, et al. Detecting remotely related proteins by their interactions and sequence similarity. Proc Natl Acad Sci U S A. 2005 May 17;102(20):7151–6.
260. Espadaler J, Eswar N, Querol E, Avilés FX, Sali A, Marti-Renom MA, et al. Prediction of enzyme function by combining sequence similarity and protein interactions. BMC Bioinformatics. 2008 May 27;9(1):249.
261. Benos PV, Bulyk ML, Stormo GD. Additivity in protein–DNA interactions: how good an approximation is it? Nucleic Acids Res. 2002 Oct 15;30(20):4442–51.
262. Pray L. Eukaryotic genome complexity. Nat Educ. 2008;1(1):96.
263. Lelli KM, Slattery M, Mann RS. Disentangling the many layers of eukaryotic transcriptional regulation. Annu Rev Genet. 2012;46:43–68.
264. Walhout AJM, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, et al. Protein Interaction Mapping in C. elegans Using Proteins Involved in Vulval Development. Science. 2000 Jan 7;287(5450):116–22.
265. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han J-DJ, et al. Annotation Transfer Between Genomes: Protein–Protein Interologs and Protein–DNA Regulogs. Genome Res. 2004 Jun 1;14(6):1107–18.
266. Wei G-H, Badis G, Berger MF, Kivioja T, Palin K, Enge M, et al. Genome□wide analysis of ETS□family DNA□binding in vitro and in vivo. EMBO J. 2010 Jul 7;29(13):2147–60.
267. Hu Z, Killion PJ, Iyer VR. Genetic reconstruction of a functional transcriptional regulatory network. Nat Genet. 2007 May;39(5):683–7.
268. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004 Sep 2;431(7004):99–104.
269. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389–402.
270. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2011 Nov 29;40(D1):D290–301.
271. Hollenhorst PC, Pietz G, Fox CA. Mechanisms controlling differential promoter-occupancy by the yeast forkhead proteins Fkh1p and Fkh2p: implications for regulating the cell cycle and differentiation. Genes Dev. 2001 Sep 15;15(18):2445–56.
272. Kafri R, Bar-Even A, Pilpel Y. Transcription control reprogramming in genetic backup circuits. Nat Genet. 2005 Mar;37(3):295–9.
273. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning [Internet]. New York, NY, USA: ACM; 2006 [cited 2014 Sep 1]. p. 233–40. Available from: http://doi.acm.org/10.1145/1143844.1143874
274. Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD. Genome editing with engineered zinc finger nucleases. Nat Rev Genet. 2010 Sep;11(9):636–46.
275. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007 Jul 1;35(suppl 2):W407–10.
276. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh C-H, et al. A Genomic Regulatory Network for Development. Science. 2002 Mar 1;295(5560):1669–78.
277. Stathopoulos A, Levine M. Whole-genome analysis of Drosophila gastrulation. Curr Opin Genet Dev. 2004 Oct;14(5):477–84.

278. Inoue T, Wang M, Ririe TO, Fernandes JS, Sternberg PW. Transcriptional network underlying Caenorhabditis elegans vulval development. Proc Natl Acad Sci U S A. 2005 Apr 5;102(14):4972–7.

279. Koide T, Hayata T, Cho KWY. Xenopus as a model system to study transcriptional regulatory networks. Proc Natl Acad Sci U S A. 2005 Apr 5;102(14):4943–8.

280. Ochoa-Espinosa A, Yucel G, Kaplan L, Pare A, Pura N, Oberstein A, et al. The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. Proc Natl Acad Sci U S A. 2005 Apr 5;102(14):4960–5.

281. Singh H, Medina KL, Pongubala JMR. Contingent gene regulatory networks and B cell fate specification. Proc Natl Acad Sci U S A. 2005 Apr 5;102(14):4949–53.

282. Lembong J, Yakoby N, Shvartsman SY. Pattern formation by dynamically interacting network motifs. Proc Natl Acad Sci. 2009 Mar 3;106(9):3213–8.

283. Zhang Y, Wong C-H, Birnbaum RY, Li G, Favaro R, Ngan CY, et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature. 2013 Dec 12;504(7479):306–10.

284. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207–10.