Chapter 7

Discussion

This thesis presents dry and wet lab techniques to elucidate the involvement of transcription factors (TFs) in the regulation of the cell cycle and myogenesis. However, the techniques described in this manuscript could be used for the study of other TFs in these and other biological processes. These two methodologies complement each other. In silico analysis provides clues into what to verify in the wet lab, which can be used as the basis for additional in silico predictions. When developing genetics and genomics to study complex cellular processes these methodologies are essential to successfully tackle the large quantities and complexities of data successfully.

7.1 In Silico Prediction of Transcription Factor Binding Sites: Past, Present, and Future

Figure 7.1: PWM evolution: The sequence affinity of TFBSs has evolved from single sequences, to PWMs, to larger and larger databases of PWMs.

Computational predictions of TF binding sites (TFBSs) have come a long way. From initial single specific sequences, such as myogenic regulatory factors binding an E-box (simply the sequence CANNTG, reviewed in Sabourin et al. 2000 (161)), the jump was made to position weight matrices (PWMs), accounting for the variation in sequences bound by specific TFs (Figure 7.1). PWMs have also accounted for combinations of TFs serving as complexes, such as the TRANSFAC PWM V$MYOGNF1_01 for Myogenin and NF1 (162). As more PWMs become available (such as in a database) we can mine sequences for multiple PWMs indicating co-regulation and competition between multiple TFs.

However, since at present there are only a few hundred PWMs and the human proteome is estimated to contain approximately 2600 proteins with DNA binding
domains (163), there is still a lot to discover. Existing methods for obtaining experimental based PWMs include analysis of ChIP-chip and ChIP-seq data. With the cost of ChIP-seq decreasing and the popularity increasing there are more and more sequences defined as TF targets, for which we can extrapolate motifs using programs like MEME (45; 46) and Gibbs samplers (47; 48; 49). The major limitation in obtaining PWMs based on ChIP-seq data is becoming the availability of good antibodies, placing the bottleneck more on the biology and less on the informatics. Hopefully in the near future we will have a greater number of PWMs, based on a larger quantity of data resulting in higher quality.

The use of a PWM to predict TFBSs has also evolved over time. In the past programs like Match (51; 55) could identify whether a single PWM matched a sequence or not, given a similarity over a given threshold. This could be done in a batch setting for multiple PWMs, but calculations were still done one TF at a time. This program is still useful today due to its speed, but has several theoretical limitations. Three limitations are using a threshold, not accounting for competition, and not considering the TF concentration in a cell. All three of these principles are accounted for in Sunflower (56). Sunflower does not report back a black or white, bound or not bound, report, but instead a probability of a TF binding a sequence. Therefore, this also models the binding affinity of a TF. The Sunflower algorithm, by nature, introduces competition between PWMs for the same nucleotide sequence. We showed improved results when looking for enrichment of TFBSs in sequences with Sunflower compared to Match in chapter 3. In addition, though the current model sets the concentration of all TFs equal by default, they can be adjusted on a TF by TF basis. In the future, models will take into account additional factors that contribute to the binding and functionality of TFs, including TF concentration, chromatin state, and methylated nucleotides. The first steps have been made in this direction: e.g. Segal et al. (81) have included TF concentrations in their model to identify TFBSs in Drosophila.

One issue, addressed in chapters 2 and 3, is the use of proper background sequences when looking for enrichment in a foreground set of sequences. One argument is that background sets should have similar properties as the foreground set to identify TFBSs properly. For example, when a foreground set has a high TA content compared to the background set there is a high likelihood that TATA binding proteins will be predicted. Therefore, TATA box proteins predicted may well be false positives. However, TFs have access to all sequences in the genome so some may argue that it is improper to make these selections. In our searches for MyoD and Myog in expression data we found that matching foreground and background promoters based on GC content improved results. Incorporating GC content into predictions for de novo motifs has also been shown to improve results (86; 87). Since CpG islands have a higher GC content by definition (10) and potentially different promoter binding behavior (8; 9) it is also appropriate to sort data on CpG content. Besides sorting on GC and CpG content, we investigated, in chapter 3 with ChIP-chip and ChIP-PET data, sorting on presumed better annotated promoters (containing a 5′ UTR in Ensembl). However, we found that this did yield better results. In the same chapter we also compared the use of random genomic sequence as background instead of promoters. Greater significance was found using random genomic regions, confirming the a priori assumption that TFs are more likely to bind sequences near genes. This is due to differences in sequence composition between genic regions and non-genic regions,
including an overall higher GC content in coding regions (164). However, with ChIP-chip promoter based foreground sets the relative enrichment of targeted TFs compared to other TFs searched for was usually not more enriched, indicating a higher number of false positives. We therefore suggest, as has been for the similar enrichment of GO terms (91; 92; 88; 89), that for identifying over-represented TFBSs in a foreground versus background set of sequences, that both foreground and background sequences have similar properties, *e.g.* GC content, CpG content, and genic or genomic basis.

### 7.2 Wet Lab Identification and Analysis of Transcription Factor Binding Sites: Past, Present, and Future

Traditional methods, like the TransFactor kit, luciferase assays, and deletion constructs, only identified one TFBS at a time. Chromatin immunoprecipitation (ChIP) permitted the isolation of all sequences bound in the cell by a given TF, but was also, at first, limited to only analyzing a small number of targets at a time by site specific PCRs. With the invention of the microarray these ChIP fragments (ChIP-chip) could be analyzed on a genomic scale, though still limited by cost and target region (*i.e.* probes on the array). Lately, the costs for genome-wide ChIP analysis has gone down and nowadays the targeting of specific regions can be avoided by the use of ChIP in conjunction with next-generation sequencing machines (ChIP-seq). ChIP-seq also requires less input material and, potentially, can identify low affinity TFBSs (35). However, ChIP-seq is still costly and requires days of preparation. With newer technologies being introduced that permit single molecule sequencing the costs and man hours to produce data will continue to decrease. Already, articles are published using the first of these machines: the Heliscope Single Molecule Sequencer by Helicos (165; 166).

Besides using ChIP-seq and ChIP-chip for direct targeting of TFBSs, other techniques can be used to infer TFBSs. As addressed in chapters 2 and 3, groups of genes with differential expression from expression studies can be used to mine for common sequence patterns indicating shared regulatory elements (*i.e.* TFs). Like ChIP-seq and ChIP-chip, expression applications previously done on a target by target scale have been upgraded to genome wide analysis. Expression, originally analyzed by simple PCRs and gels or qPCR, can be done genome wide with a microarray or in conjunction with next-generation sequencing. Next-generation sequencing for expression analysis proves more precise, reproducible, and sensitive compared to microarrays, likely due to avoiding the background issues of hybridization techniques (36). This also provides data on genes that have similar regulation, for which the regulators (*e.g.* TFs), can be searched for. Still, the location of regulatory regions of such genes has to be determined, such as promoter regions which are often loosely defined as sequences flanking the first exon of a transcript. Techniques like DeepCAGE (Cap-analysis of Gene Expression with high throughput sequencing) can refine this. With these two applications, provided in multiple cell types and conditions, we will have greater quantity and quality of regions to search for TFBSs with *in silico* analysis.

The analysis methods of data from for next-generation sequencing applications, including ChIP-seq, have moved dramatically forward. Initially the primary limitation
was time. Though sufficient for machines with longer reads, for millions of small reads
the traditional programs of BLAST (37) or BLAT (38) were not sufficient. The Eland
alignment tool provided by Illumina increased speed dramatically, but was limited to
a short read length of 32bp. Other programs were introduced that could handle
longer reads at high quality, like Rmap (40) and Maq (39), though at a longer run
time. Old algorithms were reexamined and the Burrows-Wheeler algorithm took on
new life with the current short read alignment standards: Bowtie (42) and BWA (43).
Alignments are getting faster and faster, though accuracy should be maintained. The
ability to align massive quantities of data will continue to be an important issue as
current platforms produce more data per run and future platforms are introduced
that provide even larger quantities of sequence.

As outlined in chapter 5, we find binding of regulatory proteins with discrete
peaks (most frequently around the transcription start site (TSS)), binding across a
gene with a bias for the TSS and transcript end (a so-called ”U” shape), and binding
across the whole gene (Figure 5.1C-E). Possibly, these patterns are related to the
different ways in which p300 and CBP are able to regulate transcription: the local
peaks might be associated with genes where p300/CBP bind specifically to the TFs
that regulate gene expression in contrast to the gene-wide binding where p300/CBP
regulate the expression via histone acetylation to open up the chromatin structure
facilitating transcription activation. The combination of association with TFs and
histones may account for the ”U” shaped binding.

A major problem in ChIP-seq analysis is defining a proper peak, which indicates
the target of TF binding in a sequence. These multiple binding patterns could prove
troublesome for some current peak detection algorithms. Multiple programs have been
designed for ChIP-seq analysis, SISSRs (96), QuEST (97), a pipeline by Kharchenko
and colleagues (98), and FindPeaks (99). All of these models are based on strand
biases, in which a double peak is created due to the fact that only the 5’ ends of
all DNA fragments are sequenced and for both strands (Figure 7.2A). When ChIP-
seq sample p300 (T30-2, chapter 5) is reanalyzed with GAPSS.B (chapter 4) and
we keep tags strand-separated we do see this configuration for many TFBSs (Figure
7.2B). However, this same data also shows cases of very close tags that may not follow
these models perfectly (Figure 7.2C). We also see large genomic regions of binding
(Figure 5.1D), which we speculate as histone interactions. Since these regions do
not represent peaks, but broad binding it is not clear if algorithms will detect these
properly. Distinct binding patterns, such as in Figure 7.2B, should perform well
with current prediction systems, but further evaluation should be made on broader
binding, such as in Figure 5.1D. Future programs should focus on better addressing
and identifying multiple peak-shapes, and not just one shape.

The future of ChIP-seq data production and analysis will be faster runs at higher
quality, generating more accurate data. Single molecule sequencing allows one to
sequence the DNA of a single cell. This will enable more detailed experiments with
precise expression and ChIP analysis from a single cell, not a mix of cells common to
many cultures and samples.
7.3 TFs and Disease

Many diseases are somehow associated with function and dysfunction of TFs. This work focuses on TFs involved with two biological processes, myogenesis and cell cycle (control), both of which are linked to multiple diseases (26; 18; 122; 123). In chapters 2 and 3 we identify many TFBSs over-represented in MyoD or Myogenin bound DNA fragments, indicating other TFs that also regulate these fragments. As shown in tables 3.2 and 3.3, these additional TFs already have evidence linking them to the process of myogenesis. If not already identified as disease related, these serve as ideal candidate genes for any disease study involving muscle development and regulation, of which MyoD and Myogenin are master regulators. In chapter 5 we use the same approach to identify TFBSs over-represented in p300 and CBP bound DNA fragments, serving as ideal candidate genes for CBP/p300 related diseases such as Rubinstein-Taybi Syndrome and cancer. Several of these, such as YY1, already have known relations

Figure 7.2: Peaks: ChIP-fragments (in full represented by arrows in the top panel A), with darker ends representing what is sequenced are aligned to the genome with strand specificity. The coverage of the sequences are represented by the light "peaks," shown as strand specific (top panel A) and represented as a consensus (bottom panel A). The assumption of sequence tag distribution in programs like FindPeaks, QuEST, and SISSRS is a double peak pattern (A). UCSC browser wiggle tracks (B/C) of tags plotted on the forward (F), reverse (R), or both (F&R) for a p300 ChIP-seq dataset (T30-2 from chapter 5). When we plot p300 ChIP-seq data we see these strand specific patterns for some regions (e.g. ZNF688) (B), but for other binding events the strand bias is much less prominent (e.g. SERPINE1) (C).
to cell cycle regulation.

As more ChIP-seq data series become available, we will design better PWMs to search for TFBSs. This knowledge, coupled with whole genome sequencing of patients, will lead to discovering the cause of many diseases. Besides looking at the obvious for mutations (the coding regions of these genes), mutated target TFBS should also be observed. In addition, besides looking for the loss of a TFBS, gain of TFBSs should also be identified. A gain or loss of regulation can lead to dis-regulation and disease. In the future it will be crucial to screen for this on a patient by patient basis, termed personal medicine.

7.4 The Future of Genomics

The development of several platforms that can sequence billions of base-pairs of DNA sequence in less than a week offers new solutions to existing problems, but also generate new problems. We can now look at DNA sequences genome wide without the biases of hybridization that were part of the micro-array era. This technology has moved us closer to having the means to sequence any individual’s genome at a reasonable price and speed, a step needed to truly provide personalized medication. However, billions of base-pairs in terabytes of data provides new difficulties in data analysis and storage. The bottleneck in such experiments has dramatically shifted and will continue to shift more from the wet-lab work towards bioinformatic analysis.

These large quantities of data will place demands on storage, access, and interpretation. For storage of the current next-generation sequencer data the NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra (135) has been created. However, we can expect that instead of a lab generating gigabytes of (processed) data in a week, a lab will generate terabytes in an hour. It will be constant competition between increase in data generation and the rapidly decreasing cost of storage. Besides the actual storage devices, the access speed to these devices must be considered. Even if we can keep up with the data storage, analysis will be dramatically slowed down if connections to these storages are not increased. Finally, more methods must be developed to make the data transparent. As next-generation sequencing becomes everyday practice in research groups and clinics, the tools to extrapolate the information must be made intuitive to the average biologist and clinician.

One observation from current next-generation sequencing applications, is that many unannotated parts of the genome have functional support. Therefore the ‘old’ term “junk” DNA should be appended, or even removed altogether, in the future. There is also functional support for this from other databases (Figure 6.4A), but the challenge is incorporating all of this into a usable and visible means. Evidence also exists that a majority of bases in the genome are transcribed (167). Traditionally DNA that did not code for a protein was termed junk DNA. However, it quickly became apparent that gene regulation occurred in such junk DNA, such as TFBSs. Also some TFs need space between their TFBSs and the gene promoter, so the spacer DNA can not be truly called junk. In addition, there are genes, such as regulatory non-coding RNAs, that encode an RNA, but not a protein. As we discover more and more about the genome it becomes apparent that most nucleotides serve a purpose and thus the term “junk” DNA has become meaningless.

Another fundamental process that should be addressed concerns evidence that
transcription of genes may not always occur on the traditional sense strand. More and more we find evidence for what is termed "antisense" transcription, where a gene is transcribed from the non-coding strand. Antisense transcripts have been previously reported, and in some cases even have higher expression levels than their sense counterparts (168; 36). Though not as prominent as sense strand results, in chapter 5 we do find CAGE and SAGE tags aligning to the non-coding strand. 't Hoen et al. (36) have published evidence of SAGE antisense transcription, which even indicated that in 11% of their genes antisense transcription was even more prominent than sense transcription. A human genome-wide study also presented approximately 1600 transcripts with evidence of transcription from both strands (169). We also find binding on the opposite end of the genes as expected (3' for CAGE and 5' for SAGE, CAGE example in Figure 6.1). This is also observed for CBP/p300 binding, where in addition to binding transcript starts there is transcript end binding (Figure 5.1A-B). Finding TF binding and CAGE tags at the transcript ends has also been found in other studies (138; 160). In addition, we see in the CAGE/SAGE tags aligning to the gene end on the anti-sense strand. Examples of CAGE tags aligning to the antisense strand, gene end, and both are shown in Figure 7.3. These CAGE/SAGE/ChIP-seq data that show unexpected results (antisense, gene end, or both) were regarded as artifacts in the past, but as many different methods point towards these phenomena they must be considered as a true biological process. As the increased quantity of less biased genomic data arises the process of reverse and/or antisense transcription will hopefully become more evident.

Next-generation sequencing has provided us with the means to also take the genomics field to the next level. It is becoming more cost-effective and accurate to measure gene expression and genome-wide TF binding. These methods are becoming
more quantitative, with genes expressed in exact numbers of sequences, as apposed
to previous methods (i.e. probe intensities on an array). In addition, as shown in
chapter six, we have the methods to better annotate gene structure, such as TSSs,
used in specific cells/tissues under exact conditions. This chapter demonstrates that
the current genome annotation, however impressive, is not complete. Though not
addressed in this thesis, there are also next-generation applications to identify DNA
methylation and chromatin accessibility (via ChIP-seq). The challenge of the bioin-
formatician in the near future will be combining existing and upcoming information
about gene expression, gene structure, DNA methylation, chromatin composition, TF
binding, and additional genome properties to construct a more complete model of the
entire biological processes in the genome of an animal, including man. This can even
be made more complete by combining more fields, such as proteomics, to construct a
complete picture of life and its regulation. Besides a picture of life in general, these
high throughput methods are, and will increasingly, be used to identify variation with
a single individual or population, leading to true personalized medicine and increased
effectiveness of health care.