CHAPTER 7

Discussion

Transcription factors constitute a key component of cellular networks by regulating expression of e.g. housekeeping as well as cell-type specific genes. Specific DNA-amino acid interactions allow TFs to identify their short target sequences within the many million of possible sites present in even the smallest genomes. However, in contrast to the genetic code, the regulatory interactions are highly degenerate permitting a given TF to bind not only to a single sequence but to a broad variety of sites with varying strength. This ambiguity renders the accurate prediction of target promoters for a given TF a challenging task even in the case of simple organisms such as yeast with its relatively compact genome. On the level of individual sites, TF binding is usually deduced based on generalizing from experimentally determined binding motifs (Stormo, 2000). When scanning a DNA strand each site obtains a similarity score measuring the divergence between the known motif and the sequence at hand. In part due to the desire for identifying individual motif matches classical annotation methods introduce an artificial score cutoff dividing the sequence space into binding sites and non-binding sites for the factor (Rahmann et al., 2003; Levitsky et al., 2007). Such hit-based methods thus cement a binary separation between binding and non-binding in contrast to the gradual binding behavior of transcription factors. Using discrete motif-matching approaches it has therefore been difficult to rationalize the continuous TF affinities measured in large scale ChIP experiments. Additionally, results of any subsequent analyses based on hit predictions often vary strongly depending on the choice of the employed score cutoffs (data not shown).

The first goal of this thesis was therefore to develop a new method (called TRAP) that predicts the binding affinity of a transcription factor to a DNA sequence of interest without introducing an artificial separation between binding and non-binding sites. As measure of affinity, TRAP computes the expected number of transcription factors bound to a sequence by integrating all its weak and strong binding signals. The resulting quantity $\langle N \rangle$ retains the relative binding strength of the individual sites and thus relates well to the binding data presented by ChIP-chip and PBM experiments. Correlating predicted and measured affinities for intergenic regions from yeast revealed that TFs with larger motifs tend to bind with higher affinity than TFs with short binding sites. This finding allowed to subsequently derive a generic prescription for how to set the TRAP parameters also in the absence of ChIP data. This prescription not only pertains to yeast but also to higher organisms including *Drosophila*, mouse and human. The TRAP model using the generic parameters is more successful in predicting relative binding strengths than any of the hit-based methods and also outperforms the alternative affinity based approaches that rely on a more simplified model assuming Boltzmann distributed site occupation. Importantly, while the classical log likelihood scores reside on very different scales the generic parameterization of TRAP yields predicted binding affinities that are largely comparable between different TFs without further measures. This allows not only to detect the likely target genes of a given TF but conversely also to determine which TFs regulate a given gene.

Embedding TRAP into a statistical framework that allows the robust detection of regulatory associations between TFs and groups of genes forms the second focus of this thesis. The developed method, called PASTAA, has the great advantage that gene sets do not have to be precisely defined *a priori*, instead, a ranking can be provided that reflects the association of all genes with the given input category (e.g. genes ranked according to expression level in a given tissue). Applying PASTAA to genes ranked according to tissue specificity allowed to make a number of important biological observations, for instance, that binding signals for tissue specific TFs tend to reside in proximal promoters upstream of the respective TSS and that tissue specific genes oftentimes possess a TATA box. The predictions made by PASTAA proofed to be remarkably robust against exchanging the source of expression data as well as enlarging the sequence space, especially when invoking phylogenetic footprinting. Considering sequence conservation between human and mouse not only allowed to scan larger upstream regions before the significance of the binding signals decayed but also allowed the recovery of some TF-tissue associations not detected otherwise, such as MYOD and muscle (Hewitt et al., 2008). However, some experimentally known associations like TTF1-thyroid gland were lost by restricting the sequence space only to conserved blocks. One reason for this might be technical problems such as spurious TSS annotations in one or the other species, which in turn results in an apparent lack of conserved sequence blocks. Alternatively, the binding signals for a given TF might have changed in the course of evolution. Such cases would be candidates for altered gene expression between the two species, a possibility that might be further investigated in the future.

An interesting alternative application for PASTAA is the search for coregulating TFs. While transcription is oftentimes mediated by multiple factors acting sequentially or in concert the search for TFs targeting preferentially the same sequences remains a daunting task in bioinformatics. As demonstrated for yeast and vertebrates (Sections 6.3.1 and 6.3.3), when applied to gene groups from ChIP-chip binding data PASTAA can oftentimes detect not only the TF tested in the experiment but also its coregulating factors. This applicability of PASTAA can be carried further by providing, instead of a list of genes ranked based on their association with a given data set, a gene list ranked according to predicted affinities for a given TF. The statistical test then identifies the TF with the most similar target gene ranking. To distinguish trivial results that arise due to the presence of multiple PFMs for the same TF it is necessary to take the similarity between PFMs into account and to restrict the analysis to pairs of strongly divergent TF motifs for which a significant target gene overlap has been detected. Using this approach on a number of PFMs yielded indeed interesting preliminary results. For instance, when feeding PASTAA with the entire list of 26.000 mouse genes ranked according to predicted affinities for HNF4 the method shows HNF1 as the top coregulator despite a lack of any apparent motif similarity between the corresponding PFMs. Given these findings it will be interesting to investigate whether meaningful TF-TF interactions can be detected systematically over the entire set of PFMs from TRANSFAC once a systematic procedure for removing trivial associations has been developed.

Aside from the approach for TF target gene detection taken up in this thesis, which relies on the existence of preassembled PFMs, there exists a large number of methods that aim at deriving TF motifs *de novo* (e.g. Harbison et al., 2004; Smith et al. 2005). The most common approach is hereby to select sequences suspected to be bound by the same factor and to subsequently apply a program like MEME (Timothy et al., 2006), which searches for overrepresented motifs within the supplied sequences. For gene sets stemming from ChIP-chip data more specialized approaches have been developed over the last years that apply affinity based methods to directly derive energy matrices for a given TF by optimizing the correlation between predictions and the actual *R/G* measurements (Bussemaker, Justin Kinney & Callan, Tanay). While such statistical and biophysical methods have been successfully employed for deriving matrices in yeast, for mammalian sequence sets the *de novo* motif finding approach has had limited success (Huber et al. 2006). In addition, while such methods can be very useful for obtaining binding descriptions for individual factors they do not provide a solution on how to derive accurate binding probabilities for the large number of TF for which only PFMs exist that were derived from small scale experiments. In contrast, the TRAP approach not only provides a general prescription for obtaining meaningful binding probabilities for PFMs derived from small scale experiments but may also be extended in the direction of *de novo* motif finding. In a preliminary study matrices representing all possible consensus sequences of certain length were supplied to TRAP. Using a simple ROC curve analysis the predicted affinities were then used to assess which consensus sequences rank the promoters of a given experimentally known target genes set highest. Applying this TRAP based approach to searching the 200 bp proximal promoters of groups of tissue specific genes allowed to recover the consensus sequences for a number of tissue specific TFs while MEME in contrast, was not able to identify any of the known TF motifs for the same gene sets. It thus appears possible that a further developed version of this approach, perhaps in conjunction with an efficient algorithm for developing TF motifs from consensus sites, could proof successful in the search for binding motifs even in groups of coexpressed genes from vertebrates.

The TF-DNA binding energy predictions underlying the TRAP model stem from a modified version (which adds a correction term for the genome wide base frequencies of the organism from which a TF matrix was derived) of the famous statistical mechanical selection theory developed by Berg and von Hippel (1987). These energies are in the following converted into binding probabilities via an equilibrium model that follows Fermi-Dirac statistics rather than the usually employed model assuming simplified Boltzmann statistics for binding site occupation. The accuracy of the resulting TRAP predictions depends on the validity of a number of simplifying assumptions made by either component of the physical model. Most notably, total mismatch energies are computed as the sum over independent contributions from the bases in a given site. While this is a reasonable approximation for sequences close to the consensus (Benos et al., 2002), for other sequences unrealistically large mismatch energies and in turn spuriously low binding probabilities may be predicted. Modifying the model in such a way that the sum over the mismatch contributions approaches a physically meaningful maximum might further improve the predictions. Given the noisy ChIP data available today accurately determining such parameters would likely be difficult if not impossible however.

Currently large scale experimental approaches are underway that aim at determining the binding strength between a given TF and all possible sites for the corresponding factor. Such data will greatly enhance our understanding of the mechanisms that underlie DNA-protein interactions and will allow to obtain ever more accurate biophysical as well as statistical binding models. It will be interesting to witness the impending progress made in this exciting field of bioinformatics over the next years.