Feature Based Representation and Detection of Transcription Factor Binding Sites

Rainer Pudimat, Ernst-Günther Schukat-Talamazzini, Rolf Backofen
Friedrich-Schiller-Universität Jena, Institut für Informatik
Ernst-Abbe-Platz 3, 07743 Jena
email : {rpudimat,backofen}@inf.uni-jena.de

Abstract: The prediction of transcription factor binding sites is an important problem, since it reveals information about the transcriptional regulation of genes. A commonly used representation of these sites are *position specific weight matrices* which show weak predictive power. We introduce a feature-based modelling approach, which is able to deal with various kind of biological properties of binding sites and models them via *Bayesian belief networks*. The presented results imply higher model accuracy in contrast to the PSSM approach.
Keywords: Bayesian networks, transcription factor binding sites, stochastic modelling, gene expression

1 Introduction

A fundamental challenge of recent biological research is to understand the regulation of gene expression. A gene's expression level is mainly controlled via the binding of *transcription factors* to regulatory DNA-elements (called *transcription factor binding sites*) in the upstream region of a gene [AJL+02]. Despite the quite strong sequence similarity among the binding sites of certain transcription factor, the development of highly specific and accurate computer-aided detection approaches is still an unsolved problem [LH02]. Since the relatively short sequences occurring at binding sites can show a certain degree of variability and could be present by chance anywhere in a genome without having regulatory functionality, most current solutions suffer from a high false positive rate [FH97, PBCB99].

In this paper, we consider the supervised learning of binding site motifs. Although nearby any known modelling approach has been applied, the majority of current available solutions uses *position specific weight matrices* (PSSM) [ATC+03, BOH03, KGR+03]. Each entry of such a matrix stands for the frequency of certain nucleotide (matrix rows) in certain position within the binding site motif (matrix columns) [St00].

Albeit its predominant role, PSSMs have only weak predictive power for two reasons. First, PSSMs assume statistical independence among the motif positions. Recent literature shows that this strong assumption is invalid [BJC02, BEFK03]. Second, PSSMs do not allow to employ any other binding site properties such as DNA structural properties [Gr02, PBCB99].

Our paper deals with the development of binding site models that abolish these putative

sources of weakness. We employ *Bayesian belief networks* (BNs) [Mi97], since they provide the necessary flexibility for choosing the most predictive properties of the sites. In addition, BNs overcome the second obstacle of PSSMs, in allowing to express dependencies between these properties. Note that Bayesian belief networks are commonly used in modelling problems concerning *gene expression networks* [So03, NKIM04]. Barash et al first applied them to binding site prediction in 2003 [BEFK03]. In contrast to our work their application of BNs is just an extension of the PSSM through considering dependencies between sequence positions, without modelling more complex sequence properties. Bayesian belief networks also have been used for methodical similar tasks like modelling of splice sites [CDKK00].

This paper is structured as follows: In section 2, we introduce basic principles of our modelling strategy and describe the different site properties used in the belief networks. In section 3, we perform model accuracy tests on an exemplarily data set. It is shown, how the predictiveness increases if we enrich the former approach with dependency modelling and flexible motif descriptions.

2 Methods

2.1 Model Features

The goal of learning is to detect common properties between different samples given in the data sets. For this purpose, it is common to describe these properties by a vector $F_1, \ldots, F_k$ of *features*, which can be extracted directly from the sample sequences. For PSSMs, this features are simply the nucleotides at the different positions. In our case, we have more complex features, e.g. 'being in an CpG island'. Using these features, we are also able to characterise important properties of the flanking regions like structural attributes. In addition, they are also used as a technique for parameter reduction.

Now these features $F_1, \ldots, F_k$ are modeled as discrete random variables, and the problem of learning is to estimate the joint probability distribution $P(F_1, F_2, \ldots, F_K)$ from a set of training samples. In what follows, these random variables associated to the features in our model are called *model features*. We distinguish between six main classes of features (called *feature types*) which are currently implemented: *nucleotide features*, *consensus features*, *helical parameter features*, *GC content features*, *CpG island features* and *PSSM hit features*.

Nucleotide Features represent base distributions at single motif columns, analogous to columns of a PSSM. A consensus feature is used to determine if a defined subsequence of a binding site contains a match of a given consensus sequence. Helical parameter features evaluate the mean of sequence-dependent structural dinucleotide steps over a defined sequence range according to [PPF$^+$99, EHC97]. One has the choice of 38 different conformational parameters defined on dinucleotides(for instance *helical twist* or *mayor groove width*). GC content features represent the fraction of guanine and cytosine in the neighbourhood of a site. It can be applied to approximate the overall base composition in the

environment of a site. CpG island features behave similar to GC content features but possess distinct value sets. They take the value *true* if the site is within a CpG island and *false* if the site is not. The matrix hit feature was constructed to deal with the preference of transcription factor binding sites to occur in the close neighbourhood of a co-acting transcription factor's site [La98]. Due to this, they are used to scan the flanking regions of a site for hits of a defined PSSM.

Albeit model features differ in their value range and their rules for mapping sequences to feature values, we simply assume that they are discrete functions $F_k : S \times \mathbb{N} \mapsto \mathrm{ran}(F_k)$ from the Cartesian product of the set S of DNA sequences and integers to the feature range. The additional integer input determines a reference position of the sequence. Since we cannot deal with continuous random variables, the continuous range of helical parameter features and GC content features is discretised. The interval borders for discretisation are determined by an entropy-based algorithm developed by Fayyad and Irani [FI93].

Example. Let's consider a model feature F of the type *consensus feature*. In this case let the consensus pattern be `WWWW` (`W` stands for either `A` or `T`). Further let the range to be scanned for matches be $[i+1, i+9]$. Thus, this instance of a consensus feature decides whether or not there is a `A`–`T` subsequence with flexible start point. Then given an input sequence $\boldsymbol{s} = s_1 s_2 \cdots s_L$ the value of feature F is determined as follows:

$$F(s_1 s_2 \cdots s_L, i) = \begin{cases} true & : \; s_{i+1} \cdots s_{i+9} \text{ contains a subseq matching } \texttt{WWWW} \\ false & : \; \text{otherwise} \end{cases}$$

Figure 1 gives an overview of all available model features types. Given a set of model features which forms a binding site model of certain factor, it is possible that some nucleotides contribute to more than one feature which leads to correlations between the involved features. Furthermore, it has to be pointed out that PSSMs are embedded as a special case in our modelling approach since they can be constructed out of nucleotide features for each position inside the motif.

2.2 Bayesian Belief Networks

Constructing a model for binding sites of a transcription factor requires the choice of model features $F_1, F_2, \ldots, F_K$ with maximal discriminative power. The next step is the estimation of the joint probability distribution $P(F_1, F_2, \ldots, F_K)$. If we would assume independence of the different features as done in the case of PSSMs, this joint distribution would be calculated as the product of single probabilities of the feature values, i.e., by $\prod_{i=1}^{k} P(F_i)$. But clearly, we cannot assume statistical independence between features like mentioned above. Even if there were no possibly overlapping features (like *consensus features*) in the model, recent literature reports interdependencies between the columns of binding motif [BJC02]. This implies, that the joint probability has to be combined from conditional probabilities modelling the dependencies.

However, with respect to the usual amount of training data and the exponential growing number of distribution parameters which have to be estimated, it is hardly practicable

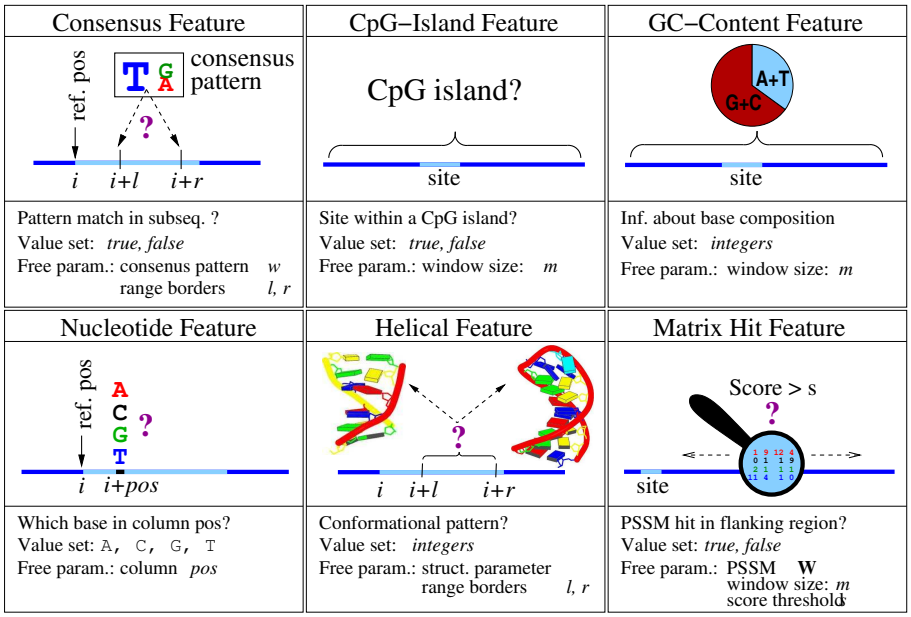

Figure 1: Model feature types

to model all putative dependencies. *Bayesian belief networks* (BN) are a good trade-off between these two extrema [BEFK03]. The motivation for the decision in favour of BNs with respect to Hidden Markov Models (HMM) is as follows. Beside the fact that they are similar in the sense of modelling dependencies, a HMM is a finite state machine. Thus, at each time point they take a state out of a state alphabet. The probability of taking a state depends on previous states of the machine. This implies an order of the states (e.g. in time series) which is not reasonable to our framework of model features whose sequence ranges could overlap. In the Bayesian belief networks approach there are no ordered states.

A BN is a pair $B = (G, P)$ where G is an annotated directed acyclic graph (DAG) whose vertices correspond to random variables out of a set $\boldsymbol{X} = \{X_1, X_2, \ldots, X_K\}$ and whose edges determine dependencies between the connected variables. Independence assumptions in the sense that each random variable is independent from all its non-parents are given implicitly by the graph structure. The parameter set P quantifies the network. It contains parameters $p_{x_k|\pi_{x_k}} = P_B(X_k = x_k \,|\, \Pi_{x_k} = \pi_{x_k})$ for each possible value x_k of random variable X_k and each assignment π_{x_k} of values to the set of parent variables Π_{x_k} [FGG97]. So, a BN B defines a unique joint probability distribution over all concerned random variables $\boldsymbol{X}$ given by

$$P_B(x_1, x_2, \ldots, x_K) \quad = \quad \prod_{k=1}^{K} P_B(x_k \,|\, \pi_{x_k}). \tag{1}$$

It is clear that our model features play the role of the random variables in the BNs. Additionally there is another variable called *class variable* C [FGG97] which can be used to learn a single BN for binding sites of different factors or to distinguish binding site subclasses of one transcription factor. Together with a class variable a BN is also called *Bayesian Classifier*. An exemplary BN trained on MEF-2 binding sites is shown in figure 2.

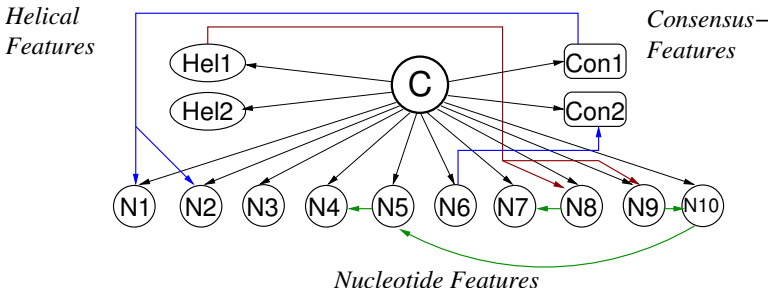

Figure 2: Bayesian belief network for MEF-2 binding sites

Learning. The learning process of Bayesian belief networks comprises both: estimating the probability distribution and the dependencies between variables (i.e. the graph structure). Since finding the best network structure of a BN given some training data is a NP-hard problem [Pe88], we restrict to three special cases of BN for which the freedom of drawing edges in the graph is more or less constrained. The simplest kind of BN considered here is the so called *Naïve Bayesian classifier* (NBC) [DH73]. There, every model feature F is dependent of the class variable C. Other dependencies are not considered. Using NBCs, one can simulate the PSSM approach. In the second class of networks employed here each model feature F depends on the class variable C and at most one other model feature F'. Beside the outgoing edges of C such networks form sets of trees. For that reason they are called *Tree-augmented networks* (TAN). There are efficient structure learning algorithms for TANs which reduces the problem to finding a maximum weighted spanning tree [CL68]. The last graph topology discussed here is quite similar to TAN with the difference that model features which have no relevance for the classification, are disconnected from the class variable and correlations between these and other features are not considered. Due to this they are called *Selective TANs* (STAN) [SP95].

To perform supervised learning BNs with transcription factor binding sites, one needs a sample set of known binding sites justified with respect to a reference position. We have chosen the reference position to be the first position according to TRANSFAC [WCF$^+$01]. In addition, we must include as much flanking regions relative to that position as the included model features demand. Each site in the sample set has to be transformed into a vector of variable assignments by applying the model feature functions. After all these vectors are presented to the network learning procedure.

Application of Trained Models. The procedure for scanning an input sequence $\boldsymbol{s} = s_1 s_2 \cdots s_L$ is quite similar to the learning process. A variable assignment vector $(f_1(l),\ f_2(l), \ldots,\ f_K(l))$ is computed for each position l of the sequence. The network returns the joint probability of each vector. Due to the fact that a model feature could use basepairs upstream of the reference position, some positions at the 5' end of a sequence cannot be evaluated (the same is the case for the 3' end).

To decide whether a sequence position is a putative binding site or not, we compare the output probability $P_B(f_1,\ f_2, \ldots,\ f_K)$ of the binding site model with the output probability $P_N(f_1,\ f_2, \ldots,\ f_K)$ of a background model, which is an equally dimensioned (according to the features which were chosen) Bayesian belief network trained on arbitrary eukaryotic promoter sequences. This is done using the common *log-odds scores*

$$S \quad = \quad \log_2 \frac{P_B(f_1,\ f_2, \ldots,\ f_K)}{P_N(f_1,\ f_2, \ldots,\ f_K)} \ . \tag{2}$$

Comparing scores of different models to decide, which is the most probable transcription factor, binding at certain position, is more difficult. One can easily reflect that models of different factors could contain distinct numbers and types of model features according to the demands to describe their binding sites. These circumstances lead to problems. First of all, the model with the higher number of features would tend to have smaller probability values. Second, it is not possible to compare probabilities produced by features of different nature (Is it better to see a `T` at position 1 with probability 0.9 or to see a helical twist above 34.5° with probability 0.75 at subsequence $s_m \cdots s_n$?). Furthermore there arises the question to which background model these probabilities should be compared to.

We start to tackle these problems by answering the last question. Let $\mathcal{U}$ be the set of all model features which occur in any model within our classification system. Then the background model is constructed by including all features $U \in \mathcal{U}$ and learning the probability distribution and network structure given the background data described above. The next step is to expand each site model (i.e. the numerator in equation 2) in a simple way to include the missing context of the background model. Let $\mathcal{F} \subset \mathcal{U}$ be the set of model features considered in a model B and $\mathcal{G} = \mathcal{U} - \mathcal{F}$ the set of features which are included in other models and in the background model N. Then the joint probability producing a value vector $\boldsymbol{u} = (\boldsymbol{f}, \boldsymbol{g})$ of the variable set $\mathcal{U} = \mathcal{F} \uplus \mathcal{G}$ is

$$P(\boldsymbol{f},\ \boldsymbol{g}) \quad = \quad P(\boldsymbol{f}) \cdot P(\boldsymbol{g} \mid \boldsymbol{f}) \ . \tag{3}$$

The first part of the product on the right is simply approximated by the joint probability $P_B(\,\cdot\,)$ of the binding site model B whereas the second part which is not modeled in B is substituted by the conditional probability of observing values $\boldsymbol{g}$ of variables in $\mathcal{G}$ given the values $\boldsymbol{f}$ of variables in $\mathcal{F}$ according to the background distribution $P_N(\,\cdot\,)$. Fortunately, efficient algorithms to approximate these conditional probabilities with Bayesian belief networks exist [LS88]. The expanded equation for computing log-odds scores is then

$$S \quad = \quad \log \frac{P(\boldsymbol{f},\ \boldsymbol{g})}{P_N(\boldsymbol{f},\ \boldsymbol{g})} \approx \log \frac{P_B(\boldsymbol{f}) \cdot P_N(\boldsymbol{g} \mid \boldsymbol{f})}{P_N(\boldsymbol{f},\ \boldsymbol{g})} \ . \tag{4}$$

The result of computing the scores in this way is that we achieve comparable scores in the sense described at the beginning of this paragraph.

3 Experiments and Results

We demonstrate the improvements of our modelling approach using a set of 26 experimental proven mammalian *MEF-2* binding sites which were taken from TRANSFAC [WCF+01], TRRD [KIA+02] and from supplementary material of Wasserman and Fickett [WF98]. The models were validated via *10-fold cross validation tests*. In each trial 90 % of the samples were used to learn a model which then was applied to the genomic sequences of length 2000 bp containing the remaining 10 % of the sample sites.

The scores assigned to each position of the test sequences were normalised to the range $[0,\ 1]$ to absolutely ensure comparable results. Positions whose scores exceeded 0.9 were counted as matches. Since additional known MEF-2 binding sites in these genomic neighbourhood of a sample were masked in the test processes and since one can assume that these well-investigated mostly muscle-specific promoter sequences don't contain unknown MEF-2 sites, only the current test sample site of each trial was treated as *true positives* (TP), other matches as *false positives* (FP).

For each sequences the distance between the score S_m of the known site to the sequence's average $\bar{S}$ (denoted by $\Delta_{\bar{S}|S_m}$) was calculated and averaged over all test cases. As a second quality measure the so-called *F-measure* was computed from the contingency tables: $F = \frac{2 \cdot r \cdot p}{r+p}$, where $p = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$ (precision) is the fraction of true positives among all matches and $r = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ (recall) is the part of real sites which were considered as matches.

We expect improvements in two dimensions, by considering dependencies and by using additional features. In the first experiment whose results can be seen in Table 1a, three models according to the three structure classes (NBC, TAN, STAN) were trained. As all of them contained only nucleotide features, thus the NBC model is equivalent to a PSSM. Both the distance $\Delta_{\bar{S}|S_m}$ and F increase when we model dependencies. In the case of TAN the improvement is stronger than in the case of STAN, the mean distance between average score and site scores even decreases in the STAN model. The graphs in Figure 3a show clearly the stronger signal of the known binding site toward the background in the TAN case.

a.)

Measure	**NBC**	**TAN**	**STAN**
$\Delta_{\bar{S}\|S_m}$	0.6226	**0.6683**	0.5789
F	0.6956	**0.8275**	0.7499

b.)

Structure	**Measure**	M_1	M_2	M_3
NBC	$\Delta_{\bar{s}\|s_m}$	0.6400	0.6924	0.7209
	F	0.7272	0.7042	0.7272
TAN	$\Delta_{\bar{s}\|s_m}$	0.6861	0.7156	0.7339
	F	0.8421	0.8421	**0.8571**

Table 1: a.) Comparison of models with different structure classes. b.) Quality measures of models with additional features.

In the second series of experiments, additional features were successively integrated. No-

tice that an algorithm for selecting the best model features is not developed yet. Hence, the choice of model features was based on visual inspection of the data. Models M_1 contain a consensus feature which evaluates whether there is a matching sequence for consensus TD starting from position 1 at the sites (D means nucleotides A, G or T). The M_2 models contain the same features plus another consensus examining whether there exists a subsequence of length 4 with consensus WWWW (nucleotides A or T. In the M_3 models, a helical feature which measures the approximated minor groove width within the site and discretises the values according to threshold 5.36 was added. The quality measure for the NBC and the TAN approach increases from M_1 to M_3 (table 1b).

In a final step we analysed the values of recall r and precision p of the PSSM-like NBC model and the best found model (TAN–M_3) with respect to different score threshold. The graph of precision is shown in Figure 3b. Clearly, the PSSM has a weaker performance compared to the multi-feature TAN model. Since there were hardly false negative errors in the test sets, the recall values are always nearby 1.

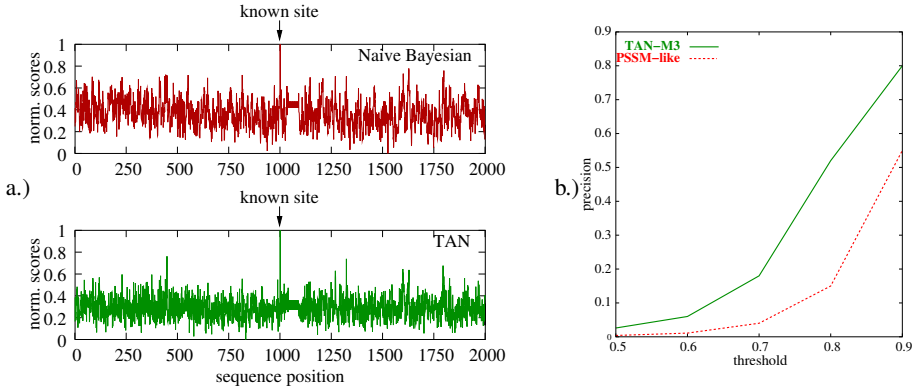

Figure 3: a.) Score series for a NBC and TAN model on the same sequence. b.) Comparison of precision values between a PSSM-like model and the multi-feature TAN model

Acknowledgment

We thank the anonymous reviewers who helped to improve the quality of this article. Furthermore we thank Ulrike Gaussmann for her helpful biological advices.

4 Conclusion

We have developed a flexible feature-based probabilistic representation of transcription factor binding sites. The predictive power of PSSMs was exceeded by both considering predictive properties of the sites (and their flanking regions), and modeling dependencies

among these properties. To represent these properties, we used Bayesian belief networks.

We have investigated different types of belief networks and compared the performance with respect to classical PSSMs. As we could show for the MEF2-transcription factor, the scores of the true sites are better distinguished from the average scores using the belief networks incorporating consensus and structural features. In addition, we have found out that the TAN (tree-augmented networks) perform better compared to STAN (selective TAN).

References

[AJL+02] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., und Walter, P.: *Molecular Biology of the Cell*. Garland Science. New York. 4. 2002.

[ATC+03] Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y., und De Moor, B.: Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Research*. 31(6):1753–64. 2003.

[BEFK03] Barash, Y., Elidan, G., Friedman, N., und Kaplan, T.: Modeling dependencies in protein-dna binding sites. In: *Proc. Seventh Annual Inter. Conf. on Computational Molecular Biology (RECOMB)*. S. 28–37. 2003.

[BJC02] Bulyk, M. L., Johnson, P. L. F., und Church, G. M.: Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research*. 30(5):1255–61. 2002.

[BOH03] Boardman, P. E., Oliver, S. G., und Hubbard, S. J.: SiteSeer: Visualisation and analysis of transcription factor binding sites in nucleotide sequences. *Nucleic Acids Research*. 31(13):3572–5. 2003.

[CDKK00] Cai, D., Delcher, A., Kao, B., und Kasif, S.: Modeling splice sites with bayes networks. *Bioinformatics*. 16(2):152–8. 2000.

[CL68] Chow, C. K. und Liu, C. N.: Approximating discrete probability distributions with dependence trees. *IEEE Trans. in Info. Theory*. 14:462–67. 1968.

[DH73] Duda, R. O. und Hart, P. E.: *Pattern Classification and Scene Analysis*. John Wiley & Sons. New York. 1973.

[EHC97] El Hassan, M. A. und Calladine, C. R.: Conformational characteristics of dna: Empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *Phil. Trans. R. Soc. Lond. A*. 355:43–100. 1997.

[FGG97] Friedman, N., Geiger, D., und Goldszmidt, M.: Bayesian network classifiers. *Machine Learning*. 29(2-3):131–163. 1997.

[FH97] Fickett, J. W. und Hatzigeorgiou, A. G.: Eukaryotic promoter recognition. *Genome Res*. 7(9):861–78. 1997.

[FI93] Fayyad, U. und Irani, K.: Multi-interval discretizationn of contnous-valued attributes for classification learning. In: *Proc. Thirtheenth International Joint Conference on Artificial Intelligence*. S. 1022–1027. Chambery, France. 1993. Morgan Kauffmann.

[Gr02] Grabe, N.: AliBaba2: context specific identification of transcription factor binding sites. *In Silico Biol*. 2(1):S1–15. 2002.

[KGR+03] Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., und Wingender, E.: MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*. 31(13):3576–9. 2003.

[KIA+02] Kolchanov, N. A., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Stepanenko, I. L., Merkulova, T. I., Pozdnyakov, M. A., Podkolodny, N. L., Naumochkin, A. N., und Romashchenko, A. G.: Transcription regulatory regions database (TRRD): its status in 2002. *Nucleic Acids Research*. 30(1):312–7. 2002.

[La98] Latchman, D. S.: *Eukaryotic transcription factors*. Academic Press. San Diego, United States. 3. 1998.

[LH02] Levy, S. und Hannenhalli, S.: Identification of transcription factor binding sites in the human genome sequence. *Mamm Genome*. 13(9):510–4. 2002.

[LS88] Lauritzen, S. L. und Spiegelhalter, D. J.: Local computations with probabilities on graphical structures and their application to expert systems. *Journal of Royal Statistical Society*. 50(2):157–224. 1988.

[Mi97] Mitchell, T. M.: *Machine Learning*. WCB/McGraw-Hill. 1997.

[NKIM04] Nariai, N., Kim, S., Imoto, S., und Miyano, S.: Using protein-protein interactions for refining gene networks estimated from microarray data by bayesian networks. In: *PSB04*. S. 336–47. 2004.

[PBCB99] Pedersen, A. G., Baldi, P., Chauvin, Y., und Brunak, S.: The biology of eukaryotic promoter prediction–a review. *Comput Chem*. 23(3-4):191–207. 1999.

[Pe88] Pearl, J.: *Probabilistic Reasoning in Intelligent Systems*. Morgan Kauffmann. 2. 1988.

[PPF+99] Ponomarenko, J. V., Ponomarenko, M. P., Frolov, A. S., Vorobyev, D. G., Overton, G. C., und Kolchanov, N. A.: Conformational and physicochemical dna features specific for transcription factor binding sites. *Bioinformatics*. 15(7-8):654–68. 1999.

[So03] Soinov, L. A.: Supervised classification for gene network reconstruction. *Biochem Soc Trans*. 31(Pt 6):1497–502. 2003.

[SP95] Singh, M. und Provan, G.: A comparison of induction algorithms for selective and non-selective bayesian classifiers. In: *Proceedings of the Twelfth International Conference on Machine Learning*. S. 497–505. 1995.

[St00] Stormo, G. D.: Dna binding sites: representation and discovery. *Bioinformatics*. 16(1):16–23. 2000.

[WCF+01] Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S., und Urbach, S.: The transfac system on gene expression regulation. *Nucleic Acids Res*. 29(1):281–3. 2001.

[WF98] Wasserman, W. W. und Fickett, J. W.: Identification of regulatory regions which confer muscle-specific gene expression. *Journal of Molecular Biology*. 278(1):167–81. 1998.