# Tanimoto's Best Barbecue:

## Discovering Regulatory Modules using Tanimoto Scores

Peter Menzel[1,2], Peter F. Stadler[2,3,4], and Axel Mosig[5,6,*]
[1]Division of Genetics and Bioinformatics, IBHV, University of Copenhagen,
Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark
[2]Bioinformatics Group, Department of Computer Science,
and Interdisciplinary Center for Bioinformatics,
University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany.
[3]Institute for Theoretical Chemistry, University of Vienna,
Währingerstrasse 17, A-1090 Vienna, Austria
[4]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico
[5]Department of Combinatorics and Geometry (DCG),
MPG/CAS Partner Institute for Computational Biology (PICB),
Shanghai Institutes for Biological Sciences (SIBS) Campus, Shanghai, China
[6]Max Planck Institute for Mathematics in the Sciences,
Inselstrasse 22, D-04103 Leipzig, Germany
[*]Corresponding Author

**Abstract:** We present a combinatorial method for discovering *cis*-regulatory modules in promoter sequences. Our approach combines "sliding window" approaches with a scoring function based on the so-called Tanimoto score. This allows to identify sets of binding sites that tend to occur preferentially in the vicinity of each other in a given set of promoter sequences belonging to co-expressed or orthologous genes. We benchmark our method on a data set derived from muscle-specific genes, demonstrating that our approach is capable of identifying modules that were identified as functional in previous studies.

## 1 Introduction

Understanding the mechanisms of transcription regulation and gene expression is still a major challenge of current genomics research. Many computational tools have been developed for processing DNA sequences to recognize the players of transcription regulation, in particular individual transcription factor binding sites (TFBS), and their functional combinations, so-called cis-regulatory modules (CRM). A major problem in this context is that TFBS are typically short while still allowing some sequence variation. Hence the TFBS patterns collected in various databases over the past decade [HWR$^+$98, SPEWL04] typically have very low specificity and, consequently, scanning large genomic regions for occurrences of these profiles necessarily leads to a large number of false positive matches. To address this issue, studying combinations of binding sites rather than occurrences of individual binding sites has attracted major attention in recent years. The biological motivation behind these approaches is that in many cases, complexes of several transcription factors are observed to regulate gene expression by binding to their respective binding sites. Consequently, these binding sites constituting CRMs tend to be located closely together on the genome, see [SvNS03, SBHLO04, PHB05, BBC$^+$06, FPB$^+$07] and the references therein. Many well-characterized examples of CRMs have can comprise more than a dozen of binding sites, typically occur several hundreds to a few thousands of base pairs upstream of the transcription start site, and have a length in the range from a few hundred to about a thousand nucleotides [FPB$^+$07].

While some insight has been gained on conservation and loss of regulatory sequences, the mechanisms underlying their evolution are largely enigmatic. In fact, sequence conservation is a suitable indicator of conserved regulatory function. Conversely, the absence of sequence conservation does not indicate loss of regulatory function [TRS05]. Recent observations [CKN$^+$00, SKC$^+$06] indicate explicitly that shuffling of conserved elements is a major mode of evolution for *cis*-regulatory elements. Such observations lay out the basis for computational procedures for discovering regulatory active elements when dealing with collections of promoters belonging to orthologous genes.

In another scenario, one is interested in understanding common regulatory elements of co-regulated genes within one species. Here, one typically drops constraints on the order of binding sites assuming that complexes of transcription factors can bind to the corresponding *cis*-regulatory module with some or all of the binding sites having changed their order or orientation. Although binding site shuffling cannot be expected to be arbitrary, there is no detailed model on eventual constraints to date. Hence, as well as for the sake of computational practicability, many discovery procedures (including the one proposed here) drop all constraints on the order and orientation of binding sites that comprise a CRM.

The outline of this paper is as follows: After dealing with related work, we sketch a generic "sliding window" approach developed in [MBPS04] that re-phrases regulatory module discovery in terms of certain set systems. On the basis of this framework, we propose an improved scoring scheme that is finally evaluated on real data sets.

## 2 Related Work

Typically, CRM discovery methods rely on binding site predictions derived from binding site databases such as TRANSFAC or JASPAR, taking into account binding site predictions in one or several promoter sequences. Kel-Margoulis *et al.* [KMIWK02] proposed a method based on identifying clusters with the property that pairwise distances between occurrences of TFBSs range within certain bounds; sets of binding sites that maximize a certain cluster score are searched by the means of a genetic algorithm. Other methods are based on probabilistic models [PHB05] or require (only sparsely and available) knowledge about interactions between transcription factors such as the algorithm presented in [SvNS03].

Among the more established methods, Sharan *et al.* proposed an approach realized in the program CREME [SBHLO04], which detects repeated occurrences of binding sites independent of their order. Recently, Blanchette *et al.* [BBC$^+$06] conducted a genome-wide survey of CRMs based on regions observed to be conserved in genome-wide alignments. They employ a scoring scheme based on co-occurrences of non-overlapping sites and unveil a large number of statistically significant putative regulatory modules that are conserved between human, mouse and rat. A statistical approach towards detecting CRMs and assessing their significance is described in [SSZ07, NC07].

Another different class of approaches identifies regulatory modules without a sliding window approach. These methods do not take proximity into account and identify sets of binding sites that are present anywhere in the promoter sequences under consideration. Such methods typically yield reasonable results if applied on short upstream regions, but they lose significance if considering sequences reaching upstream in the order of thousands of nucleotides. Among these approaches, Perco *et al.* [PKM$^+$05] devised a method based on the so-called *Tanimoto-score* as a similarity measure between sets of binding-sites. Our approach generalizes this approach to incorporate proximity between binding sites through a sliding window approach, replacing the genetic algorithm approach by enumerative methods.

## 3 Regulatory Module Discovery

### 3.1 A general framework for CRM discovery

As motivated above, a common scenario in regulatory module discovery is the following: we are given a set of promoter sequences, typically a few thousand nucleotides upstream from the respective transcription start site, of co-regulated or orthologous genes. Furthermore, we employ a database of transcription factor binding sites as the building blocks of the regulatory modules to be unveiled. In addition, we assume an upper bound $L$ on the length of a CRM, e.g., $L = 200$.

The database of TFBS is used to determine matches of potential binding sites using any of the established methods for matching position-specific weight matrices against the nucleotide sequence. Given these matches, a "sliding window" approach yields all possible sets of binding sites that co-occur within a window of length $L$. In other words, each promoter sequence is transformed into a set system, i.e., a set of sets of transcription factors (or their respective binding sites). Given the promoter regions of $K$ co-regulated (or homologous) genes, we therefore have to consider $K$ corresponding set systems rather than the sequence data themselves. Formally, given $K$ sequences $S_1, \ldots, S_K$, we denote the set system derived from $S_i$ as by $\mathcal{B}_i = \{B_{i,1}, \ldots, B_{i,\lambda_i}\}$, where $B_{i,j}$, and $\lambda_i$ denotes the number of sets contained in $\mathcal{B}_i$. With $m$ denoting the number of binding site profiles under consideration, we can represent each binding site by one integer between $1$ and $m$, so that each $B_{i,j}$ is a subset of $[1 : m]$ (writing $[a : b]$ for the closed integer interval between $a$ and $b$). The set $B_{i,j}$ contains the binding motifs of $j$th candidate interval of sequence $S_i$.

Regulatory modules can be identified with transcription factor binding motifs that tend

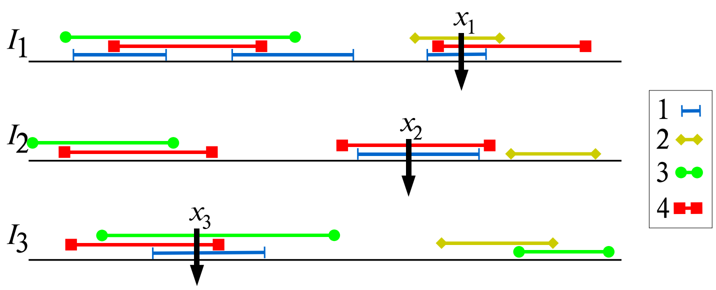

Figure 1: Illustration of the "best barbecue" approach for discovering CRMs of fixed length $L$ [MBPS04]. For each input sequence, one arrangement of colored intervals is constructed. For each transcription factor, one color is introduced; for each occurrence of the transcription factor's binding site $s$, on interval of length $L - |s|$, associated with the corresponding color, is introduced, ending at the location of the occurrence. The equivalence classes induced by the interval arrangement yield set systems, i.e., sets of colors (or integers associated with these).

to occur together within a set in several of these set systems $\mathcal{B}_i$. In the simplest case, CRMs are defined as maximal sets of TFBS that occur in all upstream regions. In the language of our set systems, we say that $A \subseteq [1 : m]$ is a $(\mathcal{B}_1, \ldots, \mathcal{B}_K)$-*barbecue* if for each $i \in [1 : K]$, there is an integer $\nu_i$ such that $A \subseteq B_{i,\nu_i}$. In [MBPS04] we develop a theory for such configurations, which we dubbed "barbecues" as they derive from a certain class of stabbing problems, as illustrated in Fig. 1. In the context of CRMs, the barbecue $A$ corresponds to a set of cooperating transcription factors (or, rather, their co-occurring binding motifs). Naturally, CRMs are identified as *"best barbecues"* which maximize the cardinality of $A$ [MBPS04].

### 3.2 Discovering CRMs using Tanimoto scores

Instead of searching for a *"Best Barbecue"*, which requires that every member of a CRM is actually present in all instances of the CRM, it seems prudent to use a "weaker" formulation. We may, for example, allow the loss of individual binding sites in a CRM, as long as the sets of binding motifs remains sufficiently similar.

The similarity between sets of binding sites is conveniently defined in terms of the *Tanimoto score*

$$\mathrm{tnm}_2(X, Y) = |X \cap Y|/|X \cup Y|, \tag{1}$$

where $|X|$ denotes the cardinality of a set $X$. By construction, the score $\mathrm{tnm}_s(X, Y)$ describes the ratio between the number of elements common in both sets and the number of all distinct elements found in both sets. The classical Tanimoto coefficient of two sets, equ.(1), always lies in the interval $[0 : 1]$ and reaches $1.0$ iff both sets are equal.

Beside this "classical" Tanimoto scoring function, we also use variants with different properties that have already been explored in the context of CRM discovery in [PKM$^+$05]:

$$\mathrm{tnm}_1(X,Y) = |X \cap Y|/|X \triangle Y|, \tag{2}$$

with $X \triangle Y$ denoting the symmetric difference between $X$ and $Y$. In equation 2 the score denotes the fraction of elements common in both sets $X$ and $Y$ and elements occurring either (exclusively) in $X$ or $Y$; as a third variant, we used

$$\mathrm{tnm}_3(X,Y) = |X \cap Y|^4/|X \cup Y|^2, \tag{3}$$

which tends to favor sets with a large intersection. Note that the score values are not necessarily contained in the interval $[0,1]$ anymore.

While these scoring functions only measure the similarity between two sets, they can be readily employed to yield a score for an arbitrary set $A \subseteq [1:m]$ of binding sites by simply calculating the sum of the Tanimoto scores over all sequences. Furthermore, we normalize the score by dividing the overall score by the number of sequences $K$. This normalization step allows to compare the results between data sets of different size. In other words, in order to assign a score to a set of binding sets $S$, we set

$$\mathrm{Tnm}(A) = \frac{1}{K} \sum_{i \in [1:K]} \max_{B \in \mathcal{B}_i} \mathrm{tnm}(A,B) \tag{4}$$

Our goal is thus to determine the set of binding sites $A$ that maximizes $\mathrm{Tnm}(A)$. In order to handle the limit the number of $2^m$ sets of binding sites to be tested, we limit our considerations to all sets $A$ with $|A| \leq \theta$, for $\theta$ typically ranging between $2$ and $5$; in general, large values of $\theta$ potentially yield higher scores, and an upper bound for the choice of $\theta$ is imposed by constraints on computation time. Limiting the value $\theta$ only slightly sacrifices a major advantage of this approach, namely being free of further parameters, such as the minimum number of sequences in which the module needs to occur (which is a common parameter for several other methods). While the authors in [PKM$^+$05] relied on a genetic algorithm for their window-less approach, we use enumerative techniques to enumerate all sets of binding sites with cardinality $\theta$ on the basis of a *revolving door* algorithm [Knu05]. In order to obtain more than just one optimal solution, we remove all binding sites occurring within the sequence intervals constituting the optimal solution, and determine the optimal solution in this reduced instance to obtain a second solution. This procedure can obviously be repeated until a desired number of solutions has been obtained.

Note that the running time of our approach is essentially exponential in either the number of overall binding sites, or at least exponential in the target module size $\theta$. This trait, shared with other CRM discovery procedures, appears to be justified in the light of the hardness result of the “best barbecue problem” discussed in [MBPS04].

## 4 Results

As a major benchmark, we applied our method to a set of promoters derived from muscle-specific genes, which was also used in [PKM$^+$05] as benchmark data. Muscle-specific

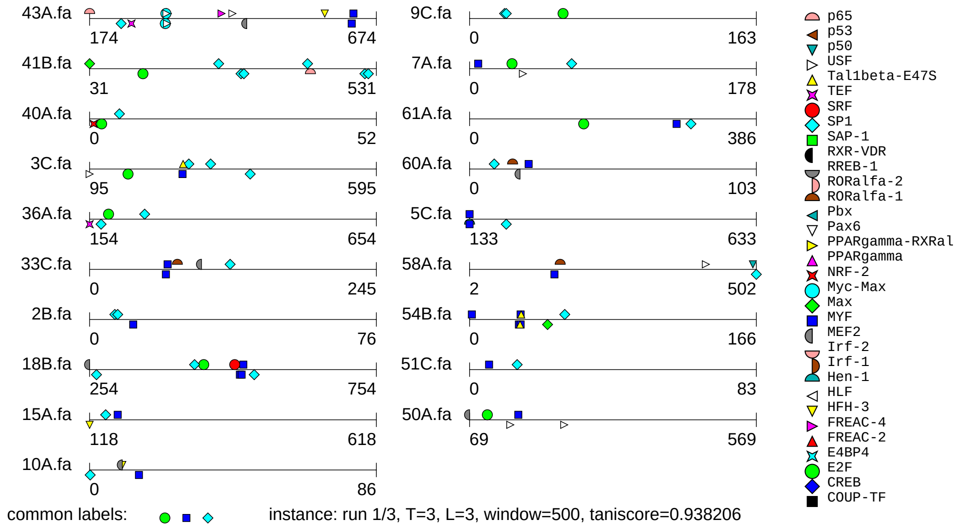

Figure 2: "Alignment" of a putative CRM of length $500$ comprising $3$ TFBSs (see *common labels* indicated in lower left part) predicted from muscle specific genes. Among the $46$ sequences under consideration, only those $19$ are shown which exhibit at least two of the three binding sites.

genes are known to be regulated by several transcription factors which interact cooperatively at multiple sites in regulatory regions. The most important factors are: *Myf*, *Mef-2*, *SRF*, *Tef-1*, and the general transcription factor *Sp-1*. Wasserman *et al.* [WF98] published a set consists of 46 DNA sequences. These promoter and enhancer regions were derived from 39 distinct human genes; the average sequence length is $333$nt, with a minimum of $61$nt and maximum of $1029$nt.

We tested the performance of the Tanimoto scores on this set of muscle-specific regulatory regions. Position weight matrices for the five known factors are taken from this publication, ensuring a high specificity of the pattern search. In addition, *Homo sapiens* PWMs with an an information content above $11$ were selected from the JASPAR CORE set [SPEWL04] and added to the the motif set. Overall, PWMs of 33 distinct transcription factor binding sites were used. No single sequence contains the combination of all five experimentally verified TFBS.

In order to test our approach, we used all three Tanimoto-style scoring functions, minimum module sizes from $2$ to $5$, minimum weights of $0.85$, $0.875$ and $0.9$ as well as module lengths of $200$, $500$ and $1000$. For each run, we retained the three best-scoring solutions.

Table 1 shows the top scoring modules using all three Tanimoto variants and a module length of $500$ with varying match thresholds $0.85$, $0.875$ and $0.9$. Note that the occurrence (occ.) columns lists the number of occurrences of the respective motif set in the highest scoring cells in each sequence, which can differ between results from the Tanimoto score variants. The three combinations of $\{$MYF, SP1$\}$, $\{$MEF2, SP1$\}$ and $\{$SP1, SRF$\}$ are the top scoring modules of size $2$, but only $\{$MEF2, SP1$\}$ and $\{$SP1,SRF$\}$ occur within all

Table 1: Highest scoring modules with match threshold of $0.85$, $0.875$, and $0.90$ and a window length of $L = 500$ in the muscle gene data set. The columns labeled # list the number of sequences in which all binding sites of the module co-occur; note that this number may eventually be $0$.

| size | $\mathtt{Tnm}_i$ | Motifs | score | # | score | # | score | # |
|---|---|---|---|---|---|---|---|---|
| | | | $0.85$ | | $0.875$ | | $0.9$ | |
| 2 | 1 | MEF2,SP1 | 0.893 | 7 | 0.811 | 5 | 0.746 | 4 |
| 2 | 2 | MYF,SP1 | 0.440 | 7 | — | | — | |
| | | MEF2,SP1 | — | | 0.403 | 5 | 0.367 | 4 |
| 2 | 3 | SP1,SRF | 0.802 | 9 | 0.704 | 8 | 0.506 | 5 |
| 3 | 1 | MEF2,MYF,SP1 | 0.893 | 4 | — | | — | |
| | | MEF2,SP1,SRF | — | | 0.766 | 2 | 0.641 | 1 |
| 3 | 2 | MEF2,MYF,SP1 | 0.386 | 4 | 0.340 | 3 | — | |
| | | MEF2,SP1,SRF | — | | — | | 0.302 | 1 |
| 3 | 3 | E2F,MYF,SP1 | 0.938 | 4 | — | | — | |
| | | MEF2,MYF,SP1 | — | | 0.711 | 3 | — | |
| | | MYF,SP1,SRF | — | | — | | 0.500 | 1 |
| 4 | 1 | E2F,MEF2,MYF,SP1 | 0.769 | 1 | — | | — | |
| | | MEF2,MYF,SP1,TEF | — | | 0.630 | 1 | — | |
| | | MEF2,MYF,SP1,SRF | — | | — | | 0.510 | 0 |
| 4 | 2 | E2F,MEF2,MYF,SP1 | 0.364 | 1 | — | | — | |
| | | MEF2,MYF,SP1,SRF | — | | 0.305 | 1 | 0.264 | 0 |
| 4 | 3 | HFH-3,MEF2,MYF,SP1 | 1.235 | 2 | — | | — | |
| | | MEF2,MYF,SP1,SRF | — | | 0.921 | 1 | — | |
| | | Irf-1,MEF2,SP1,SRF | — | | — | | 0.551 | 1 |
| 5 | 1 | E2F,MEF2,MYF,SP1,USF | 0.710 | 0 | — | | — | |
| | | MEF2,MYF,RORalfa-1,SP1,SRF | — | | 0.563 | 0 | — | |
| | | Irf-1,MEF2,MYF,SP1,SRF | — | | — | | 0.400 | 0 |
| 5 | 2 | E2F,MEF2,MYF,SP1,SRF | 0.345 | 1 | — | | — | |
| | | MEF2,MYF,SP1,SRF,TEF | — | | 0.282 | 1 | 0.234 | 0 |
| 5 | 3 | E2F,MEF2,MYF,SP1,SRF | 1.462 | 1 | — | | — | |
| | | MEF2,MYF,SP1,SRF,TEF | — | | 1.079 | 1 | — | |
| | | Irf-1,MEF2,MYF,SP1,SRF | — | | — | | 0.500 | 0 |

three match thresholds. Tanimoto variant $3$ favours $\{$SP1, SRF$\}$ over $\{$SP1, MEF2$\}$ because the former occurs four times and the latter only three times in a sequence containing the same cell (with match threshold $0.9$). These cells get an unweighted score of $4$, which results in a higher overall score when appearing four times instead of three. Generally Tanimoto variant $3$ favours candidates with most occurrences in the sequences. Top scoring modules of size $3$ are $\{$MEF2, SP1, SRF$\}$, $\{$MYF, SP1, SRF$\}$ and $\{$MEF2, MYF, SP1$\}$.

With a match threshold of $0.875$ we find a module containing all five known factors of size $5$ occurring in one sequence. Both Tanimoto variants $2$ and $3$ find this module as a Best Barbecue whereas variant $1$ finds it as third best hit.

**Comparison with Perco's GA.** The GA [PKM$^+$05] found the modules of size $2$ $\{$SP1, MEF2$\}$, $\{$SP1, MYF$\}$ and $\{$MYF, MEF2$\}$. Additionally, $\{$SP1, MEF2, MYF$\}$ was detected as best scoring module of size $3$. This module was also found by the Tanimoto bbq algorithm, as well as $\{$SP1, MEF2$\}$ and $\{$SP1, MYF$\}$. The combination of $\{$MYF, MEF$\}$ was not recognised as a high scoring module, because of the dominance of SP1, which occurs in more sequences than the other TFs. Using a match threshold of $0.85$, SP1 occurs in 33, MEF2 in 15 and MYF in 18 sequences. When removing SP1 from the set of transcripton factors, the module $\{$MEF2, MYF$\}$ is found as highest scoring module of size 2 with all three Tanimoto variants.

For further experimental exploration of our method, we refer to the more extensive investigations in the MSc thesis of the first author [Men06], which includes tests of the Best Barbecue algorithms with Tanimoto scores on artificial data sets as well as a study on $\beta$-actin related gene expression [KFW02].

## 5 Discussion

The Best-Barbecue approach translates the detection of *cis*-regulatory modules into a combinatorial "stabbing" problem, which consists in finding a maximal set of distinct TFBS that appear within an interval of given length in *each* of the input sequences [MBPS04]. While highlighting the combinatorial structure of the problem and its relationships to multiple sequence alignment problems, this strict form of the Best-Barbecue Problem (BBQ) has substantial shortcomings in practise. Most importantly, it is unlikely that every true TFBS can be detected in real sequence data: missing sequence data, binding site turnover, or simply inaccuracies in the PWMs may cause false negatives. In the strict BBQ setting, this implies that TFBS will be missing from the inferred CRMs. In fact, whenever the number of sequences is in the order of dozens (such as considered above) it becomes unlikely that the BBQ has a non-trivial solution at all, at least when realistic interval lengths are required.

In this contribution we therefore generalized the problem to using Tanimoto scores, i.e., we are attempting to stab pairwise similar sets (rather than sets that all contain the same subset) in most (rather than all) input sequences. This generalization combines the advantages of the parameter-free Tanimoto scoring scheme with the efficiency of the combinatorial stabbing approach. It increases the sensitivity of the method while retaining the specificity of sliding-window approaches on longer promoter sequences. Our computational experiments demonstrate that "Tanimoto-BBQ" is computationally feasible and produced meaningful results on data sets of practical interest.

The implementation will be publicly available with the next release of the bbq package.

## References

[BBC$^+$06] M. Blanchette, A. R. Bataille, X. Chen, C. Poitras, J. Laganiere, C. Lefebvre, G. Deblois, V. Giguere, V. Ferretti, D. Bergeron, B. Coulombe und F. Robert. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.*, 16:656–668, 2006.

[CKN$^+$00] N. A. Chuzhanova, M. Krawczak, L. A. Nemytikova, V. D. Gusev und D. N. Cooper. Promoter shuffling has occurred during the evolution of the vertebrate growth hormone gene. *Gene*, 254:9–18, 2000.

[FPB$^+$07] V. Ferretti, C. Poitras, D. Bergeron, B. Coulombe, F. Robert und M. Blanchette. `PReMod`: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res.*, 35:D122–D126, 2007.

[HWR$^+$98] T. Heinemeyer, E. Wingender, I. Reuter, H. Hermjakob, A. E. Kel, O. V. Kel, E. V. Ignatieva, E. A. Ananko, O. A. Podkolodnaya, F. A. Kolpakov, N. L. Podkolodny und N. A. Kolchanov. Databases on Transcriptional Regulation: TRANSFAC, TRRD, and COMPEL. *Nucl. Acids Res.*, 26:364–370, 1998.

[KFW02] A. Klingenhoff, K. Frech und T. Werner. Regulatory modules shared within gene classes as well as across gene classes can be detected by the same in silico approach. *In Silico Biol*, 2(1):17–26, 2002.

[KMIWK02] O. V. Kel-Margoulis, T. G. Ivanova, E. Wingender und A. E. Kel. Automatic annotation of genomic regulatory sequences by searching for composite clusters. In *Proc. Pac. Symp. Biocomput.*, Seiten 187–198, 2002.

[Knu05] D. E. Knuth. *The Art of Computer Programming, Volume 4, Fascicle 3: Generating All Combinations and Partitions.* Addison-Wesley Professional, 2005.

[MBPS04] A. Mosig, T. Bıyıkoğlu, S. J. Prohaska und P. F. Stadler. Detecting Phylogenetic Footprint Clusters by Optimizing Barbeques. *Theor. Comp. Sci.*, 2004. Preprint version: Univ. Leipzig, BIOINF 04-020, `http://www.bioinf.uni-leipzig.de/~axel/bbq.pdf`.

[Men06] P. Menzel. BBQ in Tanimoto Scores. Diplomarbeit, University of Leipzig, Germany, 2006.

[NC07] K. Noto und M. Craven. Learning probabilistic models of cis-regulatory modules that represent logical and spatial aspects. *Bioinformatics*, 23:e156–e162, 2007.

[PHB05] A. A. Philippakis, F. S. He und M. L. Bulyk. Modulefinder: a tool for computational discovery of *cis* regulatory modules. In *Proc. Pac. Symp. Biocomput.*, Seiten 519–30, 2005.

[PKM$^+$05] P. Perco, A. Kainz, G Mayer, A Lukas, R Oberbauer und B. Mayer. Detection of coregulation in differential gene expression profiles. *Biosystems*, 82:235–247, 2005.

[SBHLO04] R. Sharan, A. Ben-Hur, G. G. Loots und I. Ovcharenko. CREME: Cis-Regulatory Module Explorer for the human genome. *Nucleic Acids Res.*, 32:W253–W256, 2004.

[SKC$^+$06] R. Sanges, E. Kalmar, P. Claudiani, M. D'Amato, F. Muller und E. Stupka. Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biol.*, 7:R56, 2006.

[SPEWL04] A. Sandelin, W. A. Pär Engström, W. Wasserman und B. Lenhard. JASPAR: an open access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32:D91–D94, 2004.

[SSZ07] D. E. Schones, A. D. Smith und M. Q. Zhang. Statistical significance of cis-regulatory modules. *BMC Bioinformatics*, 8:19, 2007.

[SvNS03] S. Sinha, E. van Nimwegen und E.D. Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19:i292–i301, 2003.

[TRS05] A. Tanay, A. Regev und R. Shamir. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc. Natl. Acad. Sci. USA*, 102:7203–7208, 2005.

[WF98] W. W. Wasserman und J. W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, 278:167–181, 1998.