bbq: A Tool for Discovering Regulatory Modules using Weighted Barbeques

Axel Mosig

Bioinformatics Group, Department of Computer Science, University of Leipzig, Kreuzstrasse 7b, Leipzig, D-04103, Germany

ABSTRACT

Summary: We present a multiple-alignment-like approach for discovering clustered occurences of transcription factor binding sites, which are known to regulate the expression of genes in Eucaryotes. Our tool bbq is based on a weighted version of the so-called *best-barbeque problem*, which incorporates a p-value-like scoring scheme to detect binding site clusters that are likely to be conserved across different species due to a functional role in gene regulation.

Availability: The implementation bbq can be downloaded from http://www.bioinf.uni-leipzig.de/Software/bbq/.

Contact: Axel Mosig, Tel: ++49 341 97 16 704,
Fax: ++49 341 97 16 679, axel@bioinf.uni-leipzig.de

INTRODUCTION

Transcription in eukaryotic cells is regulated by a complex assembly of *cis*-regulatory elements that specifically bind to the DNA. The function of transcription factor binding sites (TFBSs) usually is exhibited by the synergistic action of several transcriptional activators. Here, neither the order nor the orientation of these individual binding sites is neccessarily conserved, but merely the fact that they occur clustered. Such clusters are commonly referred to as *cis-regulatory modules*. While order and orientation of the correspoding binding sites are typically conserved in homologous genes (i.e., for the same gene in different species), this is not necessarily true for genes within the same organism that are nevertheless regulated by the same combination of transcription factors. From a computatational point of view, the problem that arises therefore is to find a maximum set of short sequence fragments that occur clustered (i.e., close to each other) on large genomic sequence segments associated with different genes or different species.

Recently, a number of approaches has been proposed for the purpose of detecting *cis*-regulatory modules. Several approaches either identify statistically significant pairwise occurences of binding sites (Levy et al., 2001) or modules that consist of a known configuration of TFBSs (Wasserman and Fickett, 1998). For discovering new regulatory modules where the types of binding sites and their configuration are not known completely in advance, Sharan et al. (2004) proposed a method for identifying modules whose configuration occurs recurrently in several genomic regions.

Our approach works in the spirit of multiple alignments and seeks to find a constellation of binding sites whose clustered occurence can be detected in a number of genomic regions. Due to the multiple-alignment-like approach, our method allows to discover regulatory patterns in the upstream regions of different genes, typically of several paralog, ortholog or co-expressed genes which are suspected to share a common regulatory module.

Our implementation bbq is based on a recently developed novel algorithmic technique introduced in Mosig et al. (2004), which assigns colored intervals to binding site occurences, so that finding cis-regulatory modules can be done by solving a certain combinatorial and geometric optimization problem, the so-called best barbeque problem. As opposed to classical, typically dynamic programming based, alignment procedures, order and orientation of the binding sites' occurences can be shuffled. An implementation supporting p-value based weighting schemes and several other variants and features is publicly available.

THE BBQ SYSTEM

A typical scenario in which one seeks to identify *cis*-regulatory modules is as follows: we are given a collection of genomic upstream regions of K different genes (which are suspected to share a common regulatory module) as well as a set of m candidate binding sites. Our goal is to find a – statistically as significant as possible – subset of our m candidate binding sites that occur clustered within an interval of length L on each of the K genomes. This scenario, in fact, encompasses all input parameters of our system bbq, namely K genomic sequences, m candidate binding sites and a module length L. We will now discuss these parameters in more detail:

The *genomic sequences* are usually obtained as upstream (or possibly downstream or even intron) regions of related genes. Since in some cases, regulatory modules have been observed to be located not in the immediate vicinity of the coding region, but up to several thousand nucleotides upstream, it is a reasonable choice to consider

upstream regions of $10,000$ or even more nucleotides. The number of sequences typically ranges between two and few dozens.

One possibility of obtaining a set of *candidate binding sites* is to derive them from a position weight matrix database such as TRANSFAC (Heinemeyer et al., 1998). Alternatively, one can utilize phylogenetic footprinting tools such as `footprinter` (Blanchette and Tompa, 2003) or `tracker` (Prohaska et al., 2004), the latter one in combination with a local alignment tool such as `dialign` (Morgenstern, 1999). Applied to the K genomic sequences, one can expect that potential binding sites occur as phylogenetic footprints in these sequences. In general, there may be up to several hundred candidate binding sites, only few of which can be expected to consitute a functional module.

Cis-regulatory modules are known to have a limited length, i.e., the TFBSs that such modules consist of occur within an interval of bounded length on genomes. Here, $L = 200$ nucleotides is a common choice (Wasserman and Fickett, 1998; Sharan et al., 2004).

The output of `bbq` is a detailled description of the best weighted module that can be observed in each of the K genome sequences. Weighting is achieved through a p-value-like scoring scheme, as discussed in the following section. Beside the best weighted module, `bbq` is also capable of providing the best h modules (for an optional integer parameter h) as well as all modules whose weight exceeds a threshold t, which can also be specified as an optional input parameter. For an examplary output, see Fig. 1.

WEIGHTING SCHEMES FOR REGULATORY MODULES

A naive way of obtaining a weighting scheme for a regulatory module would be to simply count the number of binding sites contained in the cluster. However, we are interested in a statistically most significant rather than a largest cardinality regulatory module: for instance, short binding sites which occur much more frequent than long binding sites should correspondingly contribute a smaller weight to a module.

The weighting scheme implemented in `bbq` is computed from a dinucleotide-based Markov Model. We start with computing a probability $p(s,T)$ of occurence for each binding site s in each sequence T. This is achieved by considering the first order Markov model M_T resulting from the dinucleotide frequency distribution in T, which immediately yields the probability $p(s,T) := p(s|M_T)$ as the probability of binding site s being produced by M_T. Then, $w(s,T) := -\log p(s,T)$ yields a weight for an individual occurence of a binding site. Now, a module of length L occuring in T that contains the set of k binding sites $S = \{s_1,\ldots,s_k\}$ can be weighed by $w(S,T) := \sum_i w(s_i,T)$. Correspondingly, a module of length L that contains the sequences in S occuring in *all* K sequences $T_1,\ldots,T_K$ can be assigned the weight $w(S,\{T_1,\ldots,T_K\}) := \sum_j w(S,T_j)$.

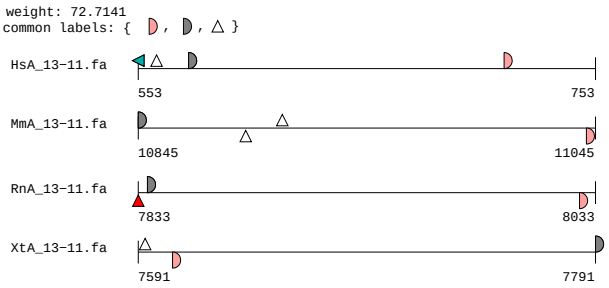

Fig. 1. A candidate regulatory module for certain Hox proteins in human, mouse, rat and xenopus obtained by applying `bbq` to intergenic regions with 41 candidate binding sites obtained by processing `tracker` generated footprints with `dialign`.

The command line tool `bbq` is capable of finding the best weighted module, based on the algorithms proposed in (Mosig et al., 2004).

Acknowledgements. Contributions of Hox cluster sequences and tracking phylogenetic footprints by Sonja J. Prohaska is gratefully acknowledged. This work was supported by the *DFG* Bioinformatics Initiative BIZ-6/1-2.

REFERENCES

Blanchette, M. and M. Tompa (2003). FootPrinter: a program designed for phylogenetic footprinting. *Nucl. Ac. Res. 31*(13), 3840–3842.

Heinemeyer, T., 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, and N. A. Kolchanov (1998). Databases on transcriptional regulation: TRANSFAC, TRRD, and COMPEL. *Nucl. Acids Res. 26*, 364–370.

Levy, S., S. Hannenhalli, and C. Workman (2001). Enrichment of regulatory signals in conseved non-coding genomic sequences. *Bioinformatics 17*(10), 871–877.

Morgenstern, B. (1999). DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics 15*, 211–218.

Mosig, A., T. Bıyıkoğlu, S. J. Prohaska, and P. F. Stadler (2004). Detecting phylogenetic footprint clusters by optimizing barbeques. submitted.

Prohaska, S., C. Fried, C. Flamm, G. Wagner, and P. Stadler (2004). Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications. *Mol. Phyl. Evol. 31*, 581–604.

Sharan, R., A. Ben-Hur, G. G. Loots, and I. Ovcharenko (2004). CREME: *Cis*-regulatory module explorer for the human genome. *Nucl. Ac. Res. 32*.

Wasserman, W. W. and J. W. Fickett (1998). Identifi cation of regulatory regions which confer muscle-specifi c gene expression. *J. Mol. Bio. 278*, 167–181.