Constructing a promoter model for antibacterial response of lung epithelial cells

Ekaterina Shelest[1], Alexander E. Kel[2] and Edgar Wingender.[1, 2]

[1]Dept. of Bioinformatics, UKG, University of Göttingen,Goldschmidtstr. 1, D-37077 Göttingen, Germany;
katya.shelest@med.uni-goettingen.de
[2] BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany;

Keywords. promoters, regulatory sequence signals, transcription factors, transcription factor binding sites, composite elements, antibacterial response.

Introduction.

Promoter model construction is a way to utilize information about coexpressed genes, which becomes more and more available with the results of microarray experiments. This growing flow of reports about coexpressed and therefore coregulated genes forces us to develop appropriate approaches and programming tools to make use of it and to go on with understanding of the underlying processes and the structure of transcriptional networks.

Binding of a bacteria to a eukaryotic cell triggers a complex network of interactions in and between both cells. *P.aeruginosa* is a pathogen that causes acute and chronic lung infections by interacting with the pulmonary epithelial cells [1,2]. We use this example for understanding the ways of triggering the response of the eukaryotic cell(s), leading us to understanding the details of the inflammatory process in general. After adhesion of *P. aeruginosa* to the epithelial cells, the response of these cells is triggered by at least two distinct agents: bacterial lipopolysaccharides [3] or bacterial pilins or flaggelins [4]. Both pathways lead to NF-kappaB activation. NF-kappaB is well-known to participate in these kinds of reactions. In this work we searched for additional transcription factors that may cooperate with NF-kappaB thus complementing and/or specifying its effects. The information about the participating transcription factors would enable us to construct a model for searching other, still not identified, target genes which are potentially involved in defensive mechanisms.

Methods

Training and Control Sets. To construct a promoter model for this purpose, we composed a training set which comprises (a) (mainly) upstream regions of genes which have been shown by microarray analysis to be affected in epithelial cells after interaction of *P. aeruginosa* [5], and (b) a smaller portion of promoters of genes the participation of which in the same response has been shown by other methods [2,4,6,7,8]. Finally the training set consisted of 24 sequences.

As the control sets we used (a) a set of 1800 5'-upstream regions, arbitrary selected from a sequence set of putative promoter regions (-350/+50) retrieved from the TRANSGENOME information resource of annotated human genome features (13000 sequences; submitted); (b) for the program using the genetic algorithm approach we used sets of specific promoters (muscle-specific, cell-cycle specific, etc); (c) a nonredundant set of the 2nd exons (retrieved from EMBL database).

Search for the Potential Transcription Factor Binding Sites. We made this search with the weight matrix approach using the Match™ tool [9]; the matrices were chosen from the library collected in TRANSFAC®[10].

Identification of Pairs. We considered all the coordinates of all potential transcription factor (TF) binding sites found by Match™ for each TF. This data was organised in tables for each sequence. Further on, we considered all possible combinations of the coordinates, thus revealing all possible pairs in the sequence.

Results

Defining the Set of Transcription Tactors (Potential Participants of the Model). In the search of the TFs potentially relevant for this type of reaction, we combined two approaches: first, biological (experimental) knowledge: some of the factors either have been reported to participate in this very process (like NF-kappaB, AP-1), or in similar processes or processes of the same type occurring in related (epithelial) cells under similar conditions (antimicrobial defence, innate immunity) (C/EBP, Elk-1). Second, we analysed the sequences for overrepresented potential binding sites applying a *genetic algorithm*. This revealed some additional TFs (Sp1, ZF5, ETF). Not all of the candidate TFs are overrepresented in the training set. No overrepresentation for such important factors, as NF-kappaB, AP-1 and C/EBP has been found. Nevertheless, these factors were included in the model, although the subsequent search will be done on the basis of overrepresentation; the reason is that not the binding sites themselves, but their combinations can be overrepresented. Finally 7 factors are included in the model: NF-kappaB, C/EBP, Sp1, AP-1, Elk-1, ETF, ZF5. The thresholds for the matrix search have been defined individually for each matrix and in such way, that at least one hit for every searched transcription factor could be found in every sequence of the training set.

Search for Combinations. To specify our model, we decided to look for combinations of potential binding sites of defined TFs. Some of the combinations could be predicted from the fact that there are well-known composite elements, like the composite elements containing NF-kappaB and C/EBP and NF-kappaB and Sp1. But there can be also other combinations, not arranged in composite elements (at least known ones). We started with a search of pairs, considering that the combinations of the higher orders can be constructed later from them. By analogy with most of the known composite elements, for which it is relevant that the pair occurs not *at* a certain distance, but *within* a certain distance, we consider the pairs occurring in segments of a certain length. We considered only if the pair (within a certain distance) appeared in the sequence at least once, not taking in account the number of pairs in each sequence. We counted the percentage of the sequences containing the considered pair, taking as 100% the number of analysed sequences (see below) in the case of the training set, and the number of sequences which contained *both constituents* of the pair in the case of the control set.

Different experiments have different level of reliability, and considering the results of these experiments one should remember about it. The microarray analysis is not absolutely reliable, so we could expect that not all of the reported genes may be relevant in the antibacterial response. Therefore, the further analysis has been made in two steps: first, we took the promoter sequences of six genes which have been reported by other methods but the microarray analysis; we took only genes independently reported by at least two different groups. We identified (see *Methods*) all the pairs present in all sequences of this "seed" group. Further on, we searched for the found pairs in the training set. We searched in the control set for the pairs found in 80-100% of the training set, and chose those which showed the most pronounced difference in the percentage (2-fold, $p<0.01$). For this quick estimation we used T-test. This search revealed one reliable pair: NF-kappaB/Elk-1 on the distance 10-22bp. Although the difference in the percentages was rather high, the percentage in the control set was high itself (0.44) and would reveal a high number of false positives during the search. Thus, we needed to find an approach which would enable us to reduce this value. The next step, therefore, was to look if there are some pairs which would be "complementary" in the sense of appearing in complementary subsets of the training set. We searched for such pairs of pairs which would cover (in complementary combination) the whole training set, but would reveal the lowest percentages in the control set. Finally we came up with 2 groups of pair combinations.

We present the approaches and the results of defining different combinations of transcription factor binding sites which can be included in a promoter model. We came up with a model which comprises (a) 2 composite elements (NF-kappaB / C/EBP and NF-kappaB / Sp1), (b) one pair of binding sites which should be present in all sequences and (c) 2 groups of pair combinations, each of which should be searched with "OR" operator. We apply this model to the search in our collection of human 5'-flanking sequences (for 13000 genes) in order to find potential target genes.

References

[1] E.DiMango, A.J.Ratner, R.Bryan, S.Tabibi and A.Prince, Activation of NF-κB by adherent *Pseudomonas aeruginosa* in normal and cystic fibrosis Respuratory epithelial cells. *J. Clin. Invest.,* 101,11:2598-2606, 1998

[2] R.S.Smith, E.R.Fedyk, T.A.Springer, N.Mukaida, B.H.Iglewski and R.P Phipps, IL-8 production in human lung fibroblasts and epithelial cells activated by the *Pseudomonas aeruginosa* autoinducer N-3-oxodododecanoyl homoserine lactone is transcriptionally regulated by NF-κB and activator protein-2. *J. immunol*., 167:366-374, 2001

[3] G.Zhang, S.Ghosh, Toll-like receptor–mediated NF-B activation: a phylogenetically conserved paradigm in innate immunity. *J. Clin. Invest*., 107, 1:13-19,2001

[4] N.McNamara, A.Khong, D .McKemy, M.Caterina, J.Boyer, D.Julius and C.Basbaum. ATP transduces signals from ASGM1, a glycolipid that functions as a bacterial receptor. *Proc. Natl. Acad. Sci.*, 98, 16: 9086-9091, 2001

[5] J.K.Ichikawa, A. Norris, M.G.Bandera, G.K.Geiss, A.B. van 't Wout, R.Bumgarner and S.Lory, Interaction of Pseudomonas aeruginosa with epithelial cells: Identification of differentially regulated genes by expression microarray analysis of human cDNAs.*Proc. Natl. Acad. Sci.*, 97, 17: 9659-9664, 2000

[6] A.Kovarik, P. J.Lu, N.Peat, J.Morris and J.Taylor-Papadimitriou, Two GC boxes (Sp1 sites) are involved in Regulation of the activity of the epithelium-specific MUC1 promoter. *J. Biol.Chem.*, 271, 30:18140-18147, 1996

[7] J.-D Li, W.Feng, M.Gallup, J-H.Kim, J.Kim, Y.Kim and C.Basbaum, Activation of NF-kB via a Src-dependent Ras-MAPK.pp90rsk pathway is required for *Pseudomonas aeruginosa*-induced micin overproductionin epithelial cells. *Proc. Natl. Acad. Sci.*, 95:5718-5723,1998

[8] M.N.Becker,G.Diamond, M.W.Verghese and S.H.Randell, CD14-dependent lipopolysaccharide-induced b-defensin-2 expression in human tracheobronchial epithelium. *J. Biol.Chem.*, 275, 38:29731-29736, 2000

[9] E.Goessling, O.V.Kel-Margoulis, A.E.Kel and E.Wingender, MATCH™ - a tool for searching transcription factor binding sites in DNA sequences. Application for the analysis of human chromosomes, in Wingender, E., Hofestädt, R., and Liebich, I. (eds.), GCB 2001, Braunschweig, Proceedings of the German Conference on Bioinformatics 2001, pages 158-161,2001

[10] E.Wingender, X.Chen, E.Fricke, R.Geffers, R.Hehl, I.Liebich, M.Krull, V.Matys., H.Michael, R.Ohnhaeuser., M.Prueß, F.Schacherer, S.Thiele and S.Urbach, The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*, 29:281-283, 2001