

BBF RFC 43: Design of Specific Mammalian Promoters by *in silico* Prediction

Tim Heinemann, Stephen Krämer, Lars Velten, Anna-Lena Kranz,
Tobias Bauer, Marti Bernardo Faura, Rainer König, Jens Keienburg,
Roland Eils and Nao Iwamoto

October 19, 2009

1 Purpose

The purpose of this RFC is to provide a) a method for the design of rational synthetic promoter sequences based on a statistical analysis about the spatial preference of transcription factor binding sites in human promoter sequences and b) further introduce standards to provide compatibility with data formats introduced in this RFC. Description of promoters generated by this method can be found at http://2009.igem.org/Team:Heidelberg/HEARTBEAT_database.

2 Relation to other BBF RFCs

BBF RFC 43 does not update or replace any earlier BBF RFC. The RFC (Request for Comments) is a new proposal for the standardized design of synthetic promoter sequences. BBF RFC 42 can be applied to synthesize promoters designed by this RFC.

3 Copyright Notice

Copyright (C) The BioBricks Foundation (2009). All Rights Reserved.

4 HEARTBEAT-DB: Starting point for the design of rational promoter sequences

Based on the assumption that transcription factors (TFs) have a spatial preference for binding to natural promoter sequences concerning the distance to the transcriptional start site (TSS) [1], the iGEM team 2009 of Heidelberg

developed HEARTBEAT (Heidelberg Artificial Transcription factor Binding site Engineering and Assembly Tool). Initially 4395 human promoter sequences 1000 bp upstream from the TSS obtained from the UCSC genome browser (source: <http://genome.ucsc.edu/>) were analysed by the bioinformatical web-tool “Promotersweep” [2]. Promotersweep is able to assign transcription factor binding sites (TFBS) to a given sequence by retrieving and combining information from three homology databases (EnsEMBL Compara, NCBI HomoloGene, DoOP database), five promoter databases (EPD, DBTSS), six sequence motive identification tools (e.g. Meme, Gibbs Motif-Sampler) and two matrix profile databases (Jaspar Core Library, Transfac Professional Library). Each TFBS motive is further classified into weak, conserved and reliable according to the quality of the assignment.

The final result of Promotersweep can be divided into general spatial information about the TFBS and the consensus sequence on the one hand and further detailed description about the associated gene on the other. This TFBS information is stored into two MySQL-tables, which helps to minimize redundant data and allows fast and efficient queries. For the statistical analysis the absolute frequency of occurrence for each TF-binding site was plotted in a histogram against the position relative to the TSS. Altogether 90 distributions with one significant peak reflecting the occurrence of a particular binding motive could be discovered. Additionally a probability density function was introduced which is more robust with respect to outliers than a normal histogram. The rescaled area under the curve between a shifting frame of 20 bases SHOULD be used as a measure for the significance of a particular TFBS occurrence. The maximum of the pdf MUST serve in the following as the position where binding sites are introduced into the rational designed promoter sequences.

5 Protocol for the standardized design of synthetic promoter sequences

The following standard protocol represented graphically in figure 1, was implemented into a graphical user interface (GUI) based on the programming language perl. The web-tool was named HEARTBEAT-GUI and is online available at http://2009.igem.org/Team:Heidelberg/HEARTBEAT_GUI.

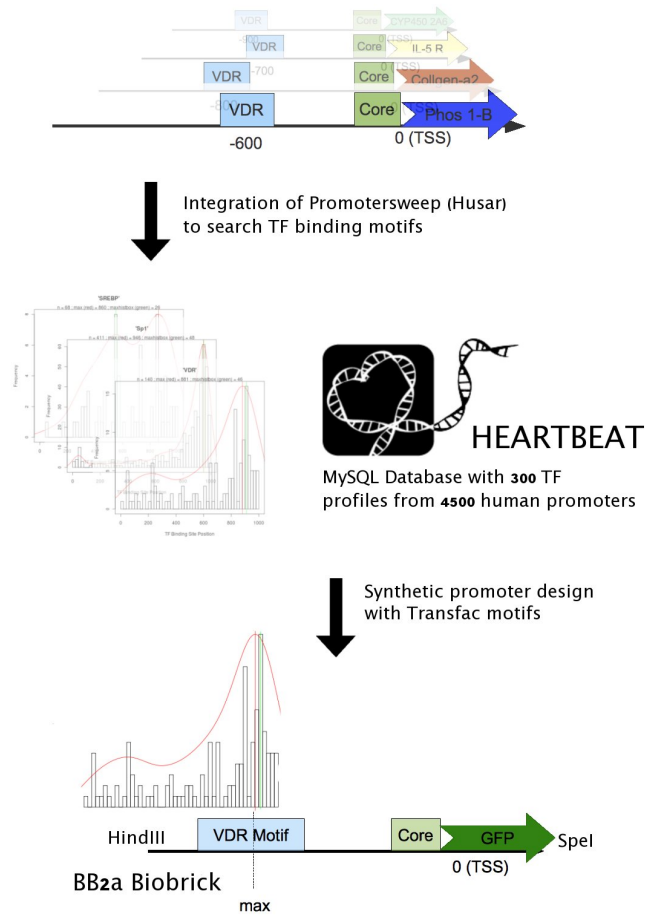


Figure 1: Graphical flow of the design of synthetic sequences

1. If the library of Transcription Factors siRNAs / constitutive TFs requested in RFC 42 is not available, a TF which is involved in a pathway which can be induced exclusively SHOULD be used to be able to evaluate the outcome of the final sequence. Potential cross-activation could lead to different transcriptional regulation.
2. CMV core promoter from Part:BBa_K203113 (separated from the proximal promoter by a *HinDIII* site) MAY be used. JeT core promoter from Part:Part:BBa_K203112 MAY also be used.
3. MUST select a corresponding consensus sequence for each chosen TF. One SHOULD use rather long sequences containing preferentially only the letters A,C,G and T from the Ambiguity Code. Consensus motives SHOULD be chosen from the Transfac database (source: <http://www.biobase-international.com/>) or the Jaspar Core Library (source: JASPAR, <http://jaspar.cgb.ki.se/>).
4. MUST create pdf from the histogram of the selected TF using the HEARTBEAT-DB.
5. MUST use maxima of the pdf (see (4)) in order to place the consensus motive with the appropriate distance in the synthetic construct.
6. MAY choose further auxiliary TFs and repeat the steps (3), (4) and (5).
7. MUST add appropriate restriction sites at the 5' and at the 3' end of the sequence to enable cloning of the construct into a reporter plasmid. *SpeI* and *HinDIII* MAY be used.
8. MUST check sequence for further restriction sites used in any BioBrick standard. (source: <http://openwetware.org/>)
9. SHOULD check sequence for further TFBS to maximize the pathway specificity of the construct. The usage of the Jaspar motive discovery tool (source: JASPAR, <http://jaspar.cgb.ki.se/>) or the “match”-tool from Transfac [3] is RECOMMENDED.

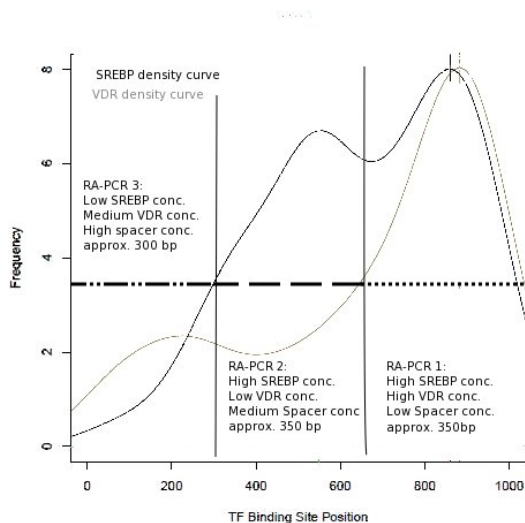


Figure 2: Probability density function of VDR and SREBP

6 M-RA-PCR

RA-PCR (RFC 42) MAY be applied for the synthesis of promoters designed by the methods outlined in this RFC. RA-PCR MUST then be modified to reflect the spatial information of the pdf-maxima. RA-PCR (RFC 42) MUST then be modified at step 3 to include a cutsite that is compatible to the cutsite from step 2 (such as *NheI*, if RFC 12 applies). Multiple RA-PCR reactions MUST then be set up, as the following example demonstrates: If a promoter regulated by multiple pathways, for example VDR (Vitamin D receptor) and SREBP (Sterol regulated element binding protein), is to be constructed, considering the density curves as obtained from the model (see figure 2) can give clues about its construction. A working VDR/SREBP promoter requires VDR and/or SREBP Response Elements (REs) in the close vicinity of the TSS (at approx. 850). It might require SREBP REs between 300 and 700, and VDR REs between 0 and 300. This distribution can be reflected by setting up 3 RA-PCRs with varying concentrations of VDR-responsive, SREBP-responsive and spacer-oligos (compare figure 2). An infinite number of RA-PCR products can be assembled and cloned in front of a core promoter.

7 Authors' Contact Information

Tim Heinemann: t.heinemann@stud.uni-heidelberg.de

Stephen Krämer: stephenkraemer@web.de

Lars Velten: velten@stud.uni-heidelberg.de

Anna-Lena Kranz: a.kranz@dkfz-heidelberg.de

Tobias Bauer: T.Bauer@Dkfz-Heidelberg.de

Rainer König: r.koenig@dkfz-heidelberg.de

Marti Bernardo Faura: marti.bernardo@bioquant.uni-heidelberg.de

Jens Keienburg: jens.keienburg@bioquant.uni-heidelberg.de

Roland Eils: r.eils@dkfz-heidelberg.de

Nao Iwamoto: n.iwamoto@dkfz-heidelberg.de

References

- [1] Daigoro K. *et al.* Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships. *Nucleic Acids Research*, pages 1–21, 2009.
- [2] del Val C. *et al.* Promotersweep: a tool for identification of transcription factor binding sites. *Theor. Chem. Acc.*, 2009. Available from World Wide Web: <http://www.springerlink.com/content/92r518nq585q716j/>.
- [3] Wingender E. *et al.* The transfac project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform.*, 2008.

For details visit <http://2009.igem.org/Team:Heidelberg/HEARTBEAT>