Artificially designed promoters
Understanding the role of spatial features and canonical binding sites in transcription
Martina Baumann, 1 Marc P. Höppner, 1 Michael Meier, 1 Jens Pontiller, 1 Wolfgang Ernst, 1 Reingard Grabherr, 1 Evan Mauceli 3 and Manfred G. Grabherr 2, 3, *

1 Department of Biotechnology; University of Natural Resources and Applied Life Sciences; Vienna, Austria; 2 Science for Life Laboratory; Department of Medical Biochemistry and Microbiology; Uppsala University; Uppsala, Sweden; 3 Broad Institute; Cambridge MA, USA

Keywords: artificial promoters, CpG content, CpG spacing, canonical binding sites, NFκB

Submitted: 10/21/11
Accepted: 10/26/11
http://dx.doi.org/10.4161/bbug.18530
*Correspondence to: Manfred G. Grabherr; Email: manfred.grabherr@imbim.uu.se

The promoter is a key element in gene transcription and regulation. We previously reported that artificial sequences rich in the dinucleotide CpG are sufficient to drive expression in vitro in mammalian cell lines, without requiring canonical binding sites for transcription factor proteins. Here, we report that introducing a promoter organization that alternates in CpGs and regions rich in A and T further increases expression strength, as well as how insertion of specific binding sites makes such sequences respond to induced levels of the transcription factor NFκB. Our findings further contribute to the mechanistic understanding of promoters, as well as how these sequences might be shaped by evolutionary pressure in living organisms.

Introduction
The transcription and regulation of genes is a complex process that, in its basic form, requires the binding of transcription factors (TF) to a promoter region, a sequence of variable length located upstream of the transcription start site of a gene. Once bound, these TFs subsequently recruit the transcriptional machinery, leading to the production of primary transcripts. Consequently, the interaction of the promoter and respective TFs initially determines the strength and timing of gene expression, and understanding the molecular basis for this interaction has therefore been subject of much discussion and research. 1,2 Previous work suggests that the binding of TFs to the promoter may be controlled by short, but defined sequence motifs. 3, 4 However, while transcription factors do appear to exhibit binding preferences for certain sequences, experimental data was able to show that they can bind to a variety of different motifs, and might even show strong affinity to more than one theme. 3, 5 This apparent promiscuity is reflected in the challenges encountered when trying to develop computational methods aiming at predicting transcription factor binding sites from nucleotide sequences, which remain prone to producing a high number of false positive candidates. 6 We may therefore ask ourselves, how exactly do TFs achieve high specificity in recognizing short binding regions out of billions of nucleotides? Is it the sequence, the constellation of critical motifs or may an even wider context be needed, meaning that short sequences in isolation carry little information to answer this question?

CpG Distribution May Determine Expression Strength Potential
We previously reported that artificial DNA sequences do not have to share homology with existing promoters to drive strong expression of a reporter gene in vitro, 7 as long as these sequences are rich in the di-nucleotide CpG, even in the absence of canonical transcription factor binding sites (TFBS). Such sequences, we found, could be bound by the generic transcription factors TFIIB and TFIIID, which would then recruit the polymerase II complex and initiate transcription. Interestingly, short binding sequences,
such as the TATA-Box, can inhibit expression if situated in certain spots in the promoter (i.e., in close proximity to the transcription start site), indicating that there is a highly localized component to TF binding, and that spacing and overall sequence arrangement might play an important role for specificity in TF/promoter interaction.²

Sequences that are rich in CpGs are found in the vast majority of active human promoters.¹⁰ The abundance of CpGs, however, appears not to be a direct predictor for expression strength potential. A total of 614 selected human sequences have been tested for in vitro activity as a promoter in vitro in several mammalian cell lines.³ The CpGs are fairly evenly spaced, with seven nucleotides being the largest distance between two consecutive instances. To determine the possible role of these motifs, we decided to alter this sequence’s organization. We replaced three or four regions of 20–25 nucleotides, roughly apart by the same distance, with nucleotide sequence devoid of CpGs (randomly chosen from the “genomic concensus”) and more abundant in A and T (Fig. 1). The ArS232d2 parental sequence and the two mutated constructs ArS232d2_3inh and ArS232d2_4inh, cloned into the reporter vector pGL3-Basic (Promega) upstream of a firefly luciferase gene, were transfected into HEK293 cells and measured for bioluminescence 24 h post transfection. Vector pGL3-Promoter (Promega), containing the SV40 promoter, served as positive control, whereas the promoterless pGL3-Basic vector was used to determine the background level. Quantitative reporter assays were done via co-transfection with the Renilla luciferase vector pRL-SV40 (Promega), used as internal standard to normalize for experimental variations. Unexpectedly, the removal of 14 CG di-nucleotides in construct ArS232d2–3inh increased expression levels markedly as compared with the construct ArS232-d2 (Fig. 2B). Deletion of another 6 CG nucleotides in construct ArS232d2–4inh decreased expression levels below the ones of ArS232-d2. By deleting said CpGs another spatial setting is creating longer intervals between adjacent CpGs. Thus, not only the total number was decreased but also the spatial arrangements. We note that the distribution of CpGs in the derived promoters more closely resembles that of existing promoters (e.g., CTAG1A, which we previously used to modulate promoter strength), suggesting that, in fact, local nucleotide changes involving CpGs can both increase and decrease activity. By changing the overall CpG content, flexibility and melting temperature also change the accessibility for DNA binding proteins and thus may cause the difference in transcription levels.

The experiment demonstrated that the activity of a promoter can be modulated by changing the number and at the same time the intervals of existing CpGs. Further mutations will serve to elucidate the optimal spatial arrangements and/or interactions of CpG islands. However it was demonstrated that for recombiant protein expression small artificial promoters are sufficient and can be modulated in terms of activity. This might further be useful since shorter promoter sequences allow for smaller plasmids to be transfected in vitro (cell line) and in vivo (e.g., in gene therapy), thereby increasing transfection efficiency.

**Figure 1.** Weak correlation between CpG richness and in vitro promoter activity. Shown is a scatterplot between CpG richness, as measured by the x-score, and in vitro promoter strength for the Stanford ENCODE Promoter data set. While the correlation is statistically significant (Spearman’s r = 0.60, p = 7.6 × 10⁻¹⁰), prediction power for individual promoter sequences based on CpG counts is very limited.

**Figure 2A** Unexpectedly, the removal of 14 CG di-nucleotides in construct ArS232-d2–3inh increased expression levels markedly as compared with the construct ArS232-d2. By deleting said CpGs another spatial setting is creating longer intervals between adjacent CpGs. Thus, not only the total number was decreased but also the spatial arrangements. We note that the distribution of CpGs in the derived promoters more closely resembles that of existing promoters (e.g., CTAG1A, which we previously used to modulate promoter strength), suggesting that, in fact, local nucleotide changes involving CpGs can both increase and decrease activity. By changing the overall CpG content, flexibility and melting temperature also change the accessibility for DNA binding proteins and thus may cause the difference in transcription levels.

The experiment demonstrated that the activity of a promoter can be modulated by changing the number and at the same time the intervals of existing CpGs. Further mutations will serve to elucidate the optimal spatial arrangements and/or interactions of CpG islands. However it was demonstrated that for recombiant protein expression small artificial promoters are sufficient and can be modulated in terms of activity. This might further be useful since shorter promoter sequences allow for smaller plasmids to be transfected in vitro (cell line) and in vivo (e.g., in gene therapy), thereby increasing transfection efficiency.

**Figure 2B**
tuned system for their collective expression and regulation. Therefore, in order for a gene to be part of such a regulatory network, its promoter needs to react to the presence of specific transcription factors to alter expression levels. In order to test whether insertion of canonical sites creates a dependency on specific transcription factors, we chose the NFκB protein, which is well studied in mammals and whose binding site consensus has been described previously. Upon insertion of five binding sites in tandem constellation, separated by CpGs, into construct ArS232, we measured in vitro activity under normal conditions as well as after induction of NFκB. HEK293 cells were transfected with the promoter constructs ArS232 and ArS232 5NFκB, inserted into the reporter vector pGL3 basic (Promega), and incubated for 12 h before induction of NFκB with recombinant human TNFα (Invitrogen) at a final concentration of 20 ng/ml. Luciferase activity was determined 6 h post induction. For standardization, the relative light units (RLU) of each promoter construct were expressed as the ratio of firefly and Renilla luciferase activity. While we note that many more experiments will be needed to explore the complexity of regulation through transcription factor binding, including determining the precise effects of spatial organization of binding sites and CpGs, initial results show that expression of construct ArS232 with binding sites increases 2.9-fold in presence of NFκB upon induction, while the expression level of the original construct, which lacks binding sites, remained roughly the same (Fig. 3). We note that we could also observe an increase in expression for the uninduced ArS232 5NFκB construct as compared with the original ArS232 construct, probably due to background levels of NFκB. Yet, it was clearly shown that induction was provided by the insertion of these transcription factor binding sites and transcription could be regulated upon external stimuli. The implications of these findings are quite remarkable, as it indicates that it should be possible to build artificial promoters that are driven by targeted TFs and can as such become part of a pathway.
Transcriptional expression and regulation is a complex process, with the promoter being only one part of the regulatory machinery within living cells (for examples and recent reviews see refs. 13–15). Our in vitro results, however, show that it is possible to make predictions for transcriptional expression based on sequence features beyond the exact nucleotide sequence. We previously demonstrated that different nucleotide sequences can act in similar ways, as long as they are rich in CpGs. Here, we touch upon how CpG richness affects expression levels leading us to the assumption that possibly also the spatial relationship of these features determines promoter activity. Further we studied how canonical binding sites embedded in CpG rich sequences can add regulation through specific transcription factors, leading us to speculate that it is the combination of both features that provides the specificity required for such regulation. On one hand, such artificially designed promoters have great potential for industrial use, e.g., for the recombinant production of antibodies, with the ability to both optimize promoter strength statically to desired levels, as well as to control them dynamically through outside stimuli. By introducing specific transcription factor binding sites and combining several of these promoter elements, it becomes possible to design artificial regulation circuits in any desired manner, e.g., to synthetically engineered monitoring systems or mechanisms that trigger cell differentiation upon certain stimuli.

On the other hand, we hypothesize that gene expression levels could be gradually modulated up or down by very local mutations in living organisms. Adaptive evolution may act on such malleable regions to adjust gene expression to optimal levels in presence of selective pressure, and that changes in the local arrangements of short sequence features allow for increasing or decreasing a promoter’s response to specific transcription factors and pathways. Likewise, near neutral changes in the affinity to specific TFs may, over time, profoundly change the expression profile of entire gene networks, giving rise to new phenotypes or even species.

Conclusions and Future Directions

References

1. Holcomb A, Tjian. Diversified transcription initiation complex expands promoter specificity and timerspecific gene expression. Gene 2003; 317: 169–201. PMID:12726286; http://dx.doi.org/10.1016/S0378-1119(03)00193-3
2. Thomas MC, Chiang CM. The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol 2006; 41:105–76. PMID:16994887; http://dx.doi.org/10.1080/10409230600648736
3. Jonsen-Gordon T, Hsu JY, Thomas PW, Kadowaki JT. The RNA polymerase II core promoter: a key component in the regulation of gene expression. Biochemie, PMID:21816204; http://dx.doi.org/10.1016/j.biochi.2011.07.024
4. Butler JR, Kadowaki J. The RNA polymerase II core promoter: a key component in the regulation of gene expression. Gene 2002; 293: 18-26. PMID:12081500; http://dx.doi.org/10.1016/S0378-1119(02)002102
5. Berger MF, Philippakis AA, Guldahl AB, Jorgensen SA, et al. Diversity and Complexity in DNA Recognition by Transcription Factors. Science 2009; 326: 726-9. PMID:19435703; http://dx.doi.org/10.1126/science.1162327
6. Xo K, Rora P, Badis G, Meissner A. A human promotersite map of candidate regulatory motif sites. Bioinformatics 2009; 25:657-64. PMID:19104035; http://dx.doi.org/10.1093/bioinformatics/btp005
7. Grabherr MG, Pontiller J, Mauceli E, Ernst W, Winzeler EA, et al. Exploring nucleotide composition to engineer promoters. PLoS ONE 2011; 6:e20136; PMID:22024501; http://dx.doi.org/10.1371/journal.pone.0020136
8. Grabherr MG, Pimentel I, Melikyan GM, Zimonjic DBB, et al. MotifMap: a human genome-wide map of candidate regulatory motif sites. Bioinformatics 2011; 27:466-74. PMID:21883893; http://dx.doi.org/10.1093/bioinformatics/btr611
9. Kim TH, Hannon GJ. The origin and evolution of microRNA. Nature 2009; 458:1010-5. PMID:19519864; http://dx.doi.org/10.1038/nature07977
10. Badis G, Berger MF, Philippakis AA, Tchkaidze I, Guldahl AB, Jorgensen SA, et al. Diversity and Complexity in DNA Recognition by Transcription Factors. Science 2009; 326: 726-9. PMID:19435703; http://dx.doi.org/10.1126/science.1162327
11. Cooper JF, Trinkin K, Anton E, Nguyen L, Mylon R. Comprehensive analysis of transcriptional promoter structure and functions in 1% of the human genome. Genome Res 2006; 16:12-29. PMID:16499906; http://dx.doi.org/10.1101/gr.426206
12. Hoffmann A, Narad G, Ghosh G. Transcriptional regulation via the NF-kappa B signaling module. Oncogene 2009; 25:7650-6; http://dx.doi.org/10.1080/10409230903305805
13. Tai R, Pang KC, Mcarthur TR, Ding A, Metcalfe D. Non-coding RNAs: regulators of disease. J Pathol 2009; 219:51-8. PMID:19472877; http://dx.doi.org/10.1002/path.2538
14. Carthew RW, Sonneville RJ. Origins and Mechanisms of miRNAs and siRNAs. Cell 2009; 136:642-53. PMID:19239886; http://dx.doi.org/10.1016/j.cell.2009.01.055
15. Miao EM, Ma N, Ambros-Ingold TM. Functional evidence of post-transcriptional regulation by pseudogenes. Biochimie, PMID:21816204; http://dx.doi.org/10.1016/j.biochi.2011.07.024