View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by IntechOpen

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

154
Countries delivered to

Our authors are among the

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

20

The Use of Functional Genomics in Synthetic Promoter Design

Michael L. Roberts
Synpromics Ltd
United Kingdom

1. Introduction

The scope of this chapter is to examine how advances in the field of Bioinformatics can be applied in the development of improved therapeutic strategies. In particular, we focus on how algorithms designed to unravel complex gene regulatory networks can then be used in the design of synthetic gene promoters that can be subsequently incorporated in novel gene transfer vectors to promote safer and more efficient expression of therapeutic genes for the treatment of various pathological conditions.

2. Development of synthetic promoters: a historical perspective

A synthetic promoter is a sequence of DNA that does not exist in nature and which has been designed to control gene expression of a target gene. *Cis*-regulatory sequences derived from naturally-occurring promoter elements are used to construct these synthetic promoters using a building block approach; which can either be carried out by rational design or by random ligation (illustrated in figure 1A). The result is a sequence of DNA composed of several distinct *cis*-regulatory elements in a completely novel orientation that can act as a promoter enhancer; typically to initiate RNA polymerase II-mediated transcription.
Construction of synthetic promoters is possible because of the modular nature of naturally-occurring gene regulatory regions. This was cleverly demonstrated by a group that used synthetic promoters to evaluate the role of the TATA box in the regulation of transcription (Mogno et al., 2010). The authors looked at the role of the TATA box in dictating the strength of gene expression. They found that the TATA box is a modular component in that its strength of binding to the RNA polymerase II complex and the resultant strength of transcription that it mediates is independent of the *cis*-regulatory element enhancers upstream. Importantly, they also found that the TATA box does not add noise to transcription, i.e. it acts as a simple amplifier without altering specificity of gene expression dictated by the upstream enhancer elements. Thus implying that any combination of *cis*-regulatory enhancers could be coupled to a TATA box and it would be the enhancers that would mediate specificity without any interference from the TATA box. The implications from this study suggest that it should be possible to construct any type of synthetic promoter that is specifically engineered to display a highly restrictive pattern of gene regulation.

Synthetic promoters have been used in the study of gene regulation for more than two decades. In one of the first examples a synthetic promoter derived from the lipoprotein gene in *E. Coli* was used to efficiently drive the expression of a number of tRNA genes (Masson et al., 1986). In the years that followed a technique was developed that enabled the mutation of prokaryotic sequences flanking the essential -10 and -35 promoter elements (illustrated in figure 1B) and thus the efficient construction of synthetic promoters for use in bacteria (Jacquet et al., 1986). This approach was successfully used to produce promoters with much higher activity compared to naturally occurring sequences and it was immediately realised that such an approach would have important applications in the biotech industry, particularly in the enhanced production of biopharmaceuticals (Trumble et al., 1992).

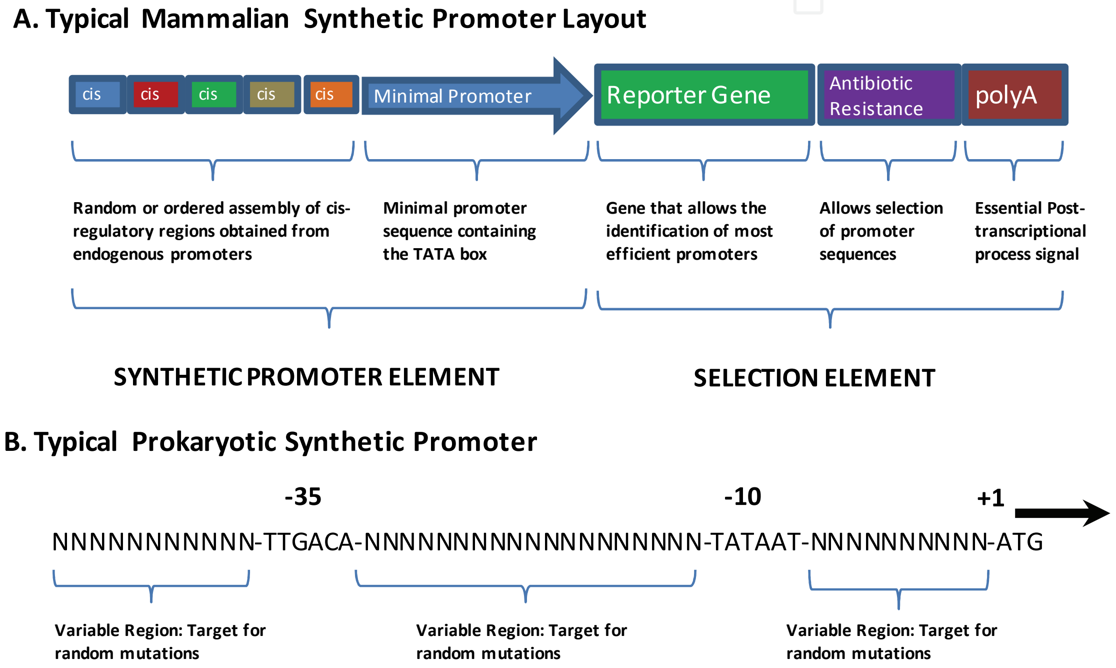

A. Typical Mammalian Synthetic Promoter Layout
cis
cis
cis
cis
cis
Minimal Promoter
Reporter Gene
Antibiotic Resistance
polyA
Random or ordered assembly of cis-regulatory regions obtained from endogenous promoters
Minimal promoter sequence containing the TATA box
Gene that allows the identification of most efficient promoters
Allows selection of promoter sequences
Essential Post-transcriptional process signal
SYNTHETIC PROMOTER ELEMENT
SELECTION ELEMENT
B. Typical Prokaryotic Synthetic Promoter
-35
-10
+1
NNNNNNNNNNN-TTGACA-NNNNNNNNNNNNNNNNNNN-TATAAT-NNNNNNNNNN-ATG
Variable Region: Target for random mutations
Variable Region: Target for random mutations
Variable Region: Target for random mutations

Fig. 1. Typical synthetic promoter layouts for prokaryotes and eukaryotes

Most of these studies were initially undertaken with a view to establish the important structural features of prokaryotic or eukaryotic promoters so that essential elements could be identified. In one example, the role of the Tat protein in the regulation of HIV gene expression was studied using synthetic promoters (Kamine et al., 1991). In this study a series of minimal promoters containing Sp1- binding sites and a TATA box were constructed and analysed to see if the Tat protein from HIV could activate them. The results demonstrated that Tat could only activate the synthetic promoters containing Sp1 sites and not promoters with the TATA box alone. The observations enabled the authors to propose that *in vivo* the Tat protein is brought to the promoter site by TAR RNA and then interacts with Sp1 to drive gene expression. In recent years more sophisticated studies using synthetic promoters have been undertaken to evaluate the important factors driving transcription factor binding to their corresponding *cis*-regulatory elements (Gertz et al., 2009a) and to thermodynamically model *trans*-factor and *cis*-element interactions (Gertz et al., 2009b).

As alluded to above, it was soon realised that synthetic promoter technology had direct implications in the improvement of the efficiency of gene expression. Indeed, one of the most widely used eukaryotic promoters employed for research purposes today is actually a synthetic promoter. The steroid-inducible Glucocorticoid Receptor Element (GRE) is a naturally occurring sequence that regulates the expression of a plethora of genes that are responsive to glucocorticoids. In a relatively early study several of these elements were linked together in order to construct a promoter with enhanced responsiveness to these steroids (Mader et al., 1993). This study detailed the construction of a 5 x GRE synthetic promoter linked to the Adenovirus type 2 major late promoter TATA region that displayed 50-fold more expression levels in response to steroid hormones when compared to the natural promoter sequence. This synthetic promoter is now a widely used constituent of a number of reporter constructs adopted in a variety of different research applications.

Finally, synthetic promoters have also been used in prokaryotic systems to reveal that regulation of gene expression follows boolean logic (Kinkhabwala et al., 2008). In this prototypical study the authors found that two transcription repressors generate a NOR logic; i.e. a OR b (on OR off), while one repressor plus one activator determines an ANDN logic; i.e. a AND NOT b (on AND NOT off). This idea was later expanded on to demonstrate that various combinations of synthetic promoters could combine to generate 12 out of 16 boolean logic terms (Hunziker et al., 2010). Most interestingly the results from these studies demonstrated that if a promoter does not follow a specific logic it is more likely to be leaky, in that it will drive gene expression under conditions where it is not expected to.

In this chapter we describe the evolution of synthetic promoter technology, its application in the development of improved tissue-specific promoters and its potential use for the development of effective disease-specific gene regulators; thus enabling the development of safer and more effective gene therapies.

3. Recent advances in the design of the synthetic promoter

In recent years some efforts have been made to construct synthetic promoters for tissue specific transcription based on the linking of short oligonucleotide promoter and enhancer elements in a random (Li et al., 1999; Edelman et al., 2000) or ordered (Chow et al., 1997; Ramon et al., 2010) fashion.

In what can be described as one of the first attempts to rationally design a tissue-specific synthetic promoter, Chow et al. describe the rearrangement of the cytokeratin K18 locus to construct a promoter mediating a highly restrictive pattern of gene expression in the lung epithelium (Chow et al., 1997). In this study the authors describe the generation of transgenic mice with this construct and demonstrate expression only in the lung. They also generated CMV (Cytomegalovirus) and SV40 (Sarcoma Virus 40) promoter based constructs and found lack of specificity and no expression in the lung epithelia. This study had important implications for researchers developing lung-based gene therapies, i.e. if CMV, one of the most widely used promoters, could not regulate gene expression in the lung epithelia then it is necessary to identify (or develop) new promoters that can efficiently regulate gene expression in this location. Indeed, it is now becoming increasingly apparent that traditional virus-derived promoters like CMV and RSV (Rous Sarcoma Virus) will have limited application in the development of modern gene therapeutics.

The random assembly of *cis*-regulatory elements has shown particular success as a means to develop synthetic promoters. In one such approach, which aimed to identify synthetic promoters for muscle-specific expression, duplex oligonucleotides from the binding sites of muscle-specific and non-specific transcription factors were randomly ligated and cloned upstream of a minimal muscle promoter driving luciferase (Li et al. 1999). Approximately

1000 plasmid clones were individually tested by transient transfection into muscle cells and luciferase activity was determined in 96-well format by luminometry. By this approach several highly active and muscle specific promoters were identified that displayed comparable strength to the most commonly used viral promoters such as CMV.

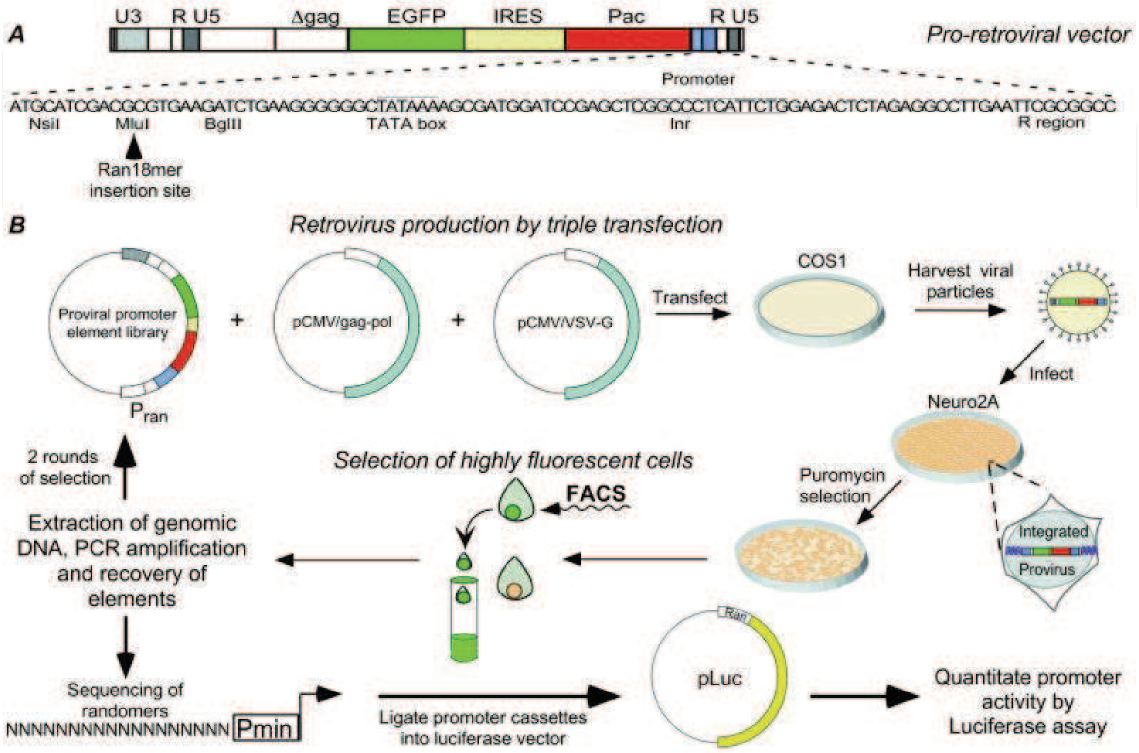

A
U3
R U5
Δgag
EGFP
IRES
Pac
R U5
Pro-retroviral vector
Promoter
ATGCATCGACGCGTGAAGATCTGAAGGGGGGCTATAAAAGCGATGGATCCGAGCTCGGCCCTCATTCTGGAGACTCTAGAGGCCTTGAATTCGCGGCC
NsiI
MluI
BglII
TATA box
Inr
R region
Ran18mer insertion site
B
Retrovirus production by triple transfection
Proviral promoter element library
P_{ran}
+
pCMV/gag-pol
+
pCMV/VSV-G
Transfect
COS1
Harvest viral particles
Infect
Neuro2A
Puromycin selection
Integrated Provirus
Selection of highly fluorescent cells
FACS
2 rounds of selection
Extraction of genomic DNA, PCR amplification and recovery of elements
Sequencing of randomers
NNNNNNNNNNNNNNNNNN Pmin
Ligate promoter cassettes into luciferase vector
Ran
pLuc
Quantitate promoter activity by Luciferase assay

Fig. 2. Typical procedure for generation of synthetic mammalian promoters (reproduced from PNAS, Vol. 97, No. 7, pp. 3038-3043 copyright (c) 2000 by the National Academy of Sciences, USA)

Retroviral vectors have also been used to screen for synthetic promoters in eukaryotic cells (Edelman et al., 2000). This study was the first description of a retroviral library approach using antibiotic resistance and FACS selection to isolate promoter sequences (illustrated in figure 2). The libraries generated using random oligonucleotides in an effort to identify new sequences as well as examining the effects of combinations of known elements and for uncovering new transcriptional regulatory elements. After preparing a Ran18 promoter library comprises random 18mer oligonucleotides, the authors analysed the sequences of the generated synthetic promoters by searching for known transcription factor binding motifs. They found that the highest promoter activities were associated with an increased number of known motifs. They examined eight of the best known motifs; AP2, CEBp, gre, ebox, ets, creb, ap1 AND sP1/maz. Interestingly, several of the promoter sequences contained none of these motifs and the author's looked for new transcription factors.

In a similar effort employed to examine one million clones, Sutton and co-workers adopted the FACS screening approach based on the establishment of a lentiviral vector-based library (Dai et al., 2004). In this study duplex oligonucleotides from binding sites of endothelial cell-specific and non-specific transcription factors were cloned in a random manner upstream of a minimal promoter driving expression of eGFP in a HIV self-inactivating expression vector. A pool of one million clones was then transfected into endothelial cells and the highest expressers were selected by FACS sorting. Synthetic promoters were then rescued from stable transfectants by PCR from the genomic DNA where the HIV vectors had integrated. The results from this study also demonstrated the possibility of isolating several highly active endothelial cell-specific synthetic promoter elements from a random screen.

Synthetic promoters active only in the liver have also been developed (Lemken et al., 2005). In this study transcriptional units from ApoB and OTC genes were used in a controlled, non-random construction procedure to generate a series of multimeric synthetic promoters. Specifically, 2x, 4x, 8x and 16x repeats of the ApoB and OTC promoter elements were ligated together and promoter activity analysed. The results indicated that the promoter based on 4xApoB elements gave the optimal levels of gene expression and that 8x and 16x elements gave reduced levels of expression, thus demonstrating the limitations of simply ligating known promoter elements together in a repeat fashion to achieve enhanced expression.

When adopting this type of methodology in the design of synthetic tissue-specific promoters it is important to use well-designed duplex oligonucleotides. For example, each element has to be spaced in such a way that the regulatory elements appear on the same side of the DNA helix when reassembled, relevant minimal promoter elements have to be employed so that the screen produces promoters capable of expressing efficiently only in the tissue of interest and there must be some sort of mechanism, such as the addition of Sp1 sites, for the protection against promoter silencing through methylation.

In addition to tissue-specific promoters, cell-type synthetic promoters have also been developed. In one study, researchers designed a synthetic promoter to be active in nonadrenergic (NA) neurones (Hwang et al., 2001). They authors randomly ligated *cis*-regulatory elements that were identified from the human dopamine beta-hydroxylase (hDBH) gene and constructed promoters with up to 50-fold higher activity than the original promoter. Specifically, two elements from the promoter were used to generate a multimeric synthetic promoter; PRS1 and PRS2 which are bound to by the Phox2a transcription factor. The results demonstrated that the PRS2 was responsible for higher levels of gene expression as it had higher affinity to Phox-2a. It was also found that eight copies of PRS2 in the same orientation yielded maximum activity.

In a similar type of study a synthetic promoter was constructed that was specifically active in myeloid cells (He et al., 2006). The promoter comprised myeloid-specific elements for PU.1, C/EBPalpha, AML-1 and myeloid-associated elements for Sp1 and AP-1, which were randomly inserted upstream of the p47-phox minimal promoter. Synthetic promoters constructed showed very high activity. Haematopoietic Stem Cells (HSC) were initially transduced then the expression in differentiated cells was examined; only myeloid cells were found to express the reporter construct. To test therapeutic applicability of these promoters apoE-/- mice were transplanted with HSC transduced with a lentiviral vector expressing apoE from CMV and synthetic promoters. Even though transduced cells containing CMV and synthetic promoters both corrected the artherosclerotic phenotype, the cells derived from lentiviral vectors harbouring the synthetic promoter did so with less variability. Thus highlighting the improved safety features when using synthetic promoters for gene therapy applications.

In addition to tissue- and cell type-specific constitutive promoters, inducible synthetic promoters can also be constructed. One group describe a synthetic promoter constructed by placing the EPO enhancer region upstream of the SV40 promoter. The result is a strong promoter that is active only under ischaemic conditions. The authors tested this promoter by developing Neural Stem Cells (NSC) responsive to hypoxia and proposed that this system could be used to deliver therapeutic stem cells to treat ischaemic events. The authors were able to demonstrate that transplantation of NSC modified with a hypoxia-sensitive synthetic promoter resulted in specific expression of the luciferase reporter gene in response to ischaemic events *in vivo* (Liu et al., 2010).

4. Applications of synthetic promoter technology

Synthetic promoters have direct applications in large-scale industrial processes where enzymatic pathways are used in the production of biological and chemical-based products (reviewed in Hammer et al., 2006). One of the most important limitations in industrial-scale processes that synthetic promoter technology addresses is the inherent genetic instability in synthetically engineered biological systems. For instance, in prokaryotic organisms designed to express two or more enzymes, mutations will invariably arise in very few generations resulting in the termination of gene expression. This is because there is the lack of evolutionary pressure keeping all the components intact. The result is that mutations accrue over generations resulting in the deactivation of the circuit. Homologous recombination in natural promoters driving high levels of gene expression is the main reason why this circuitry fails (Sleight et al., 2010). Therefore, the use of synthetic promoters in these systems should serve to lower gene expression to result in more genetic stability, allow the avoidance of repeat sequences to prevent recombination and allow the use inducible promoters (a feature that also reduces genetic instability). In summary, the use of synthetic promoter technology in complex genetically engineered synthetic organisms expressing a variety of components should serve to increase genetic stability and improve the efficiency of the processes that the components control.

One interesting therapeutic application for synthetic promoter technology that has been described is the generation of a class of replication-competent viruses that enable tumour cell-specific killing by specifically replicating in cancer cells. In this study a replication competent retrovirus was developed to selectively kill tumour cells (Logg et al., 2002). The authors added a level of transcriptional targeting by incorporating the prostate-specific probasin (PB) promoter into the retroviral LTR and designed more efficient synthetic promoters based on the PB promoter to increase the efficiency of retroviral replication in prostate cancer cells. The result was a retrovirus that could efficiency transduce and replicate only in cancer cells. This is an attractive therapeutic strategy for the treatment of cancer, as tumour virotherapy has actually been examined as a potential therapeutic strategy for several decades.

Synthetic promoters that are active only in cycling endothelial cells would be another attractive tool for the development of cancer gene therapies. The rationale being that by targeting new blood vessels growing into tumours we would be able to develop a cancer gene therapy that could cut off supply of nutrients to the growing cancer. In a study that adopted this approach the cdc6 gene promoter was identified as a candidate promoter active only in cycling cells and was coupled to the endothelin enhancer element to construct a promoter active in dividing endothelial cells (Szymanski et al., 2006). Four endothelin elements conjugated to the cdc6 promoter gave the optimal results *in vitro*. When introduced into tumour models *in vivo*, the synthetic promoter was more efficient at driving gene expression in cancerous tissues, when compared to a CMV promoter.

Perhaps one of the most impressive applications of synthetic promoter technology thus far was the development of a liver-specific promoter that could be used to essentially cure diabetes in a transgenic mouse model (Han et al., 2010). In this study a synthetic promoter active in liver cells in response to insulin was constructed. The authors designed 3-, 6- & 9-element promoters based on random combinations of HNF-1, E/EBP and GIRE *cis*-elements. In the 3-element promoters all 27 combinations of the three were tested and the highest activity promoters were used to generate the 6-element promoter and so on. Using this technique promoters with activity up to 25% of CMV were identified. Finally, the optimal promoter was chosen depending on its responsiveness to glucose. This promoter showed highest specificity to liver cells and in response to Glucose and yielded expression levels 21% that of CMV. Adenoviral vectors containing this promoter driving expression of insulin were injected into a mouse diabetic model. Injection with the highest dose of virus resulted in protection against hyperglycaemia for 50 days. Importantly, injection with adenovirus expressing insulin from a CMV promoter resulted in death of the animals due to hypoglycaemia, thus illustrating the importance of regulated expression in gene therapy. Importantly, the results from this study excellently illustrated why the clever design of synthetic promoters controlling restricted gene expression will be essential in the development of safe gene therapy.

Synthetic promoters are increasingly being used in gene therapy type of studies. In one recent study their potential application to the gene therapy of Chronic Granulomatous Disease (X-CGD; an X-linked disorder resulting from mutations in gp91-phox, whose activity in myeloid cells is important in mounting an effective immune response) was examined (Santilli et al., 2011). The authors cite a clinical trial using a retroviral vector, which was successful at correcting the phenotype, but expression was short-lived due to promoter inactivation. In order to address this issue a chimeric promoter was constructed that was a fusion of Cathepsin G and c-Fes minimal promoter sequences, which are specifically active in cells of the myeloid lineage. This promoter was used to drive the expression of gp91-phox in myeloid cells in mice using a SIN lentiviral vector and the results show effective restricted expression to monocytes and subsequent introduction of gp91 results in high levels of expression in target cells and restoration of wild type phenotype *in vitro*. X-CGD cells were then transduced with the lentiviral vector and grafted into a mouse model of CGD. The vector was able to sustain long-term expression of gp91-phox, resulting in levels of expression that could correct the phenotype. Expression was specifically seen in granulocytes and monocytes, and not B- and T-cells.

These studies serve to highlight the potential application of synthetic promoter technology in gene therapy. They particularly highlight the importance of achieving cell-type specific gene expression and address the common issue of promoter shutdown that is seen when using stronger viral promoters like those derived from the CMV and RSV. If gene therapy is to be a success in the clinic it will be imperative to develop promoters that are highly specific and which display a restrictive and predictable expression profile. Thus, synthetic promoter technology represents the ideal solution to achieve this goal and its use is likely to become an increasingly popular approach adopted by researchers developing gene therapeutics.

5. Bioinformatic tools and synthetic promoter development

We first described how functional genomics experimentation and bioinformatics tools could be applied in the design of synthetic promoters for therapeutic and diagnostic applications several years ago (Roberts, 2007). Since then a number of scientists have also realised that this approach can be broadly applied across the biotech industry (Venter et al., 2007). In this section we discuss some of the tools that we use to analyse data obtained from large-scale gene expression analyses, which is subsequently used in the smart design of synthetic promoters conveying highly-specific regulation of gene expression.

To design a synthetic promoter it is essential to identify an appropriate number of *cis*-regulatory elements that can specifically bind to the *trans* factors that enhance gene transcription. This is where the importance of a number of bioinformatic algorithms becomes apparent. Over the past several years a number of databases and programs have been developed in order to identify transcription factor biding sites (TFBSs) on a variety of genomes. Below we introduce the most extensively used resources and discuss their application to the design of synthetic promoters, we pay particular attention to the identification of transcription networks active in cancer and how this information can be used to design cancer-specific promoters that can be used in the design of safer and more effective tumour-targeted gene therapies.

There is now a growing trend for researchers to analyse microarray data in terms of 'gene modules' instead of the presentation of differentially regulated gene lists. By grouping genes into functionally related modules it is possible to identify subtle changes in gene expression that may be biologically (if not statistically significantly) important, to more easily interpret molecular pathways that mediate a particular response and to compare many different microarray experiments from different disease states in an effort to uncover the commonalities and differences in multiple clinical conditions. Therefore, we are moving into a new era of functional genomics, where the large datasets generated by the evaluation of global gene expression studies can be more fully interpreted by improvements in computational methods. The advances in functional genomics made in recent years have resulted in the identification of many more *cis*-regulatory elements that can be directly related to the increased transcription of specific genes. Indeed, the ability to use bioinformatics to unravel complex transcriptional pathways active in diseased cells can actually serve to facilitate the process of choosing suitable *cis*-elements that can be used to design synthetic promoters specifically active in complex pathologies such as cancer.

In cancer the changes in the gene expression profile are often the result of alterations in the cell's transcription machinery induced by aberrant activation of signalling pathways that control growth, proliferation and migration. Such changes result in the activation of transcription regulatory networks that are not found in normal cells and provide us with an opportunity to design synthetic promoters that should only be active in cancerous cells. If microarray technology is to truly result in the design of tailored therapies to individual cancers or even patients, as has been heralded, it is important that the functional genomics methodology that was designed for the identification of signalling and transcription networks be applied to the design of cancer-specific promoters so that effective gene therapeutic strategies can be formulated (Roberts & Kottaridis, 2007). The development of bioinformatics algorithms for the analysis of microarray datasets has largely been applied in order to unravel the transcription networks operative under different disease and environmental conditions. To this date there has been no effort to use this type of approach to design synthetic promoters that are operative only under these certain disease or environmental conditions.

The regulation of gene expression in eukaryotes is highly complex and often occurs through the coordinated action of multiple transcription factors. The use of *trans*-factor combinations in the control of gene expression allows a cell to employ a relatively small number of transcription factors in the regulation of disparate biological processes. As discussed herein, a number of tools have been developed that allow us to utilise microarray data to identify novel *cis*-regulatory elements. It is also possible to use this information to decipher the transcriptional networks that are active in cells under different environmental conditions. In yeast, the importance of the combinatorial nature of transcriptional regulation was established by specifically examining clusters of upregulated genes for the presence of combinations of *cis*-elements. By examining microarray data from yeast exposed to a variety of conditions the authors were able to construct a network of transcription revealing the functional associations between different regulatory elements. This approach resulted in the identification of key motifs with many interactions, suggesting that some factors serve as facilitator proteins assisting their gene-specific partners in their function. The idea that a core number of transcription factors mediate such a vast array of biological responses by adopting multiple configurations implies that it may be possible to hijack the transcriptional programs that have gone awry in multifactorial diseases in an effort to develop disease-specific regulatory elements. For instance, the meta-analyses of cancer datasets has permitted the identification of gene modules, allowing for the reduction of complex cancer signatures to small numbers of activated transcription programs and even to the identification of common programs that are active in most types of cancer. This type of analysis can also help to identify specific transcription factors whose deregulation plays a key role in tumour development. In one such study, the importance of aberrant E2F activity in cancer was reaffirmed during a search for the regulatory programs linking transcription factors to the target genes found upregulated in specific cancer types (Rhodes et al., 2005). It was shown that E2F target genes were disproportionately upregulated in more than half of the gene expression profiles examined, which were obtained from a multitude of different cancer types. It was thus proposed that integrative bioinformatics analyses have the potential to generate new hypotheses about cancer progression.

Different bioinformatics tools, examples of which are given in table 1, may be used to screen for *cis*-regulatory elements. In general, such tools function by comparing gene expression profiles between differentially regulated genes and examining upstream sequences, available through genome sequence resources. For the phylogenetic footprinting tools, the untranslated regions of specific genes are compared between species and the most highly conserved sequences are returned and proposed to be potential *cis*-elements. A combination of all available approaches may be employed in order to identify regulatory sequences that predominate in the profile of specific cell or tissue types. The most common sequences identified are then used as the building blocks employed in the design of synthetic promoters.

The ability to use gene expression data to identify gene modules, which mediate specific responses to environmental stimuli (or to a diseased state) and to correlate their regulation to the *cis*-regulatory elements present upstream of the genes in each module, has transformed the way in which we interpret microarray data. For instance, by using the modular approach it is possible to examine whether particular gene modules are active in a variety of different cancers, or whether individual cancers require the function of unique gene modules. This has allowed us to look for transcriptional commonalities between different cancers, which should aid in the design of widely applicable anti-cancer therapeutic strategies. In one early study, gene expression data from 1975 microarrays, spanning 22 different cancers was used to identify gene modules that were activated or deactivated in specific types of cancer (Segal et al., 2004). Using this approach the authors found that a bone osteoblastic module was active in a number of cancers whose primary metastatic site is known to be the bone. Thus, a common mechanism of bone metastasis between varieties of different cancers was identified, which could be targeted in the development of novel anticancer therapies.

It is also possible to identify the higher-level regulator that controls the expression of the genes in each module (Segal et al., 2003). Examination of the upstream regulatory sequences of each gene in a module may reveal the presence of common *cis*-regulatory elements that are known to be the target of the module's regulator. Therefore, by identifying specific regulatory proteins that control the activation of gene modules in different cancers, it should be possible to extrapolate the important *cis*-elements that mediate transcription in the transformed cell. Thereby, allowing us to design and construct novel tumour-specific promoters based on the most active *cis*-regulatory elements in a number of tumour-specific gene modules. The ability to identify specific transcriptional elements in the human genome that control the expression of functionally related genes is transforming the application of functional genomics. Until recently the interpretation of data from microarray analysis has been limited to the identification of genes whose function may be important in a single pathway or response. How this related to global changes in the cellular phenotype had been largely ignored, as the necessary tools to examine this simply did not exist. With the advancement of bioinformatics we are now in a position to utilise all the data that is obtained from large-scale gene expression analysis and combine it with knowledge of the completed sequence of the human genome and with transcription factor, gene ontology and molecular function databases, thereby more fully utilising the large datasets that are generated by global gene expression studies.

For nearly two decades scientists have been compiling databases that catalogue the *trans*-factors and *cis*-elements that are responsible for gene regulation (Wingender et al., 1988). This has primarily been done in an effort to elucidate the various transcription programs that are activated in response to different biological stimuli in a range of organisms. The result is the emergence of useful tools that can be used to identify transcription factors and their corresponding *cis*-regulatory sequences that are useful in the design of synthetic promoters. In the remaining part of this chapter we briefly discuss each resource, indicating the unique aspect of its functionality.

TRANSFAC is perhaps the most comprehensive TFBS database available and indexes transcription factors and their target sequences based solely on experimental data (Matys et al., 2003). It is maintained as a relational database, from which public releases are made available via the web. The release consists of six flat files. At the core of the database is the interaction of transcription factors (FACTOR) with their DNA-binding sites (SITE) through which they regulate their target genes (GENE). Apart from genomic sites, 'artificial' sites which are synthesized in the laboratory without any known connection to a gene, e.g., random oligonucleotides, and IUPAC consensus sequences are also stored in the SITE table. Sites must be experimentally proven for their inclusion in the database. Experimental evidence for the interaction with a factor is given in the SITE entry in form of the method that was used (gel shift, footprinting analysis, etc.) and the cell from which the factor was derived (factor source). The latter contains a link to the respective entry in the CELL table. On the basis of those, method and cell, a quality value is given to describe the 'confidence' with which an observed DNA- binding activity could be assigned to a specific factor. From a collection of binding sites for a factor nucleotide weight matrices are derived (MATRIX). These matrices are used by the tool Match™ to find potential binding sites in uncharacterized sequences, while the program Patch™ uses the single site sequences, which are stored in the SITE table. According to their DNA-binding domain transcription factors are assigned to a certain class (CLASS). In addition to the more 'planar' CLASS table a hierarchical factor classification system is also used.

TRANSCompel® originates from COMPEL, and functions to emphasize the key role of specific interactions between transcription factors binding to their target *cis*-regulatory elements; whilst providing specific features of gene regulation in a particular cellular content (Kel-Margoulis et al., 2002). Information about the structure of known *trans* factor and *cis* sequence interactions, and specific gene regulation achieved through these interactions, is extremely useful for promoter prediction. In the TRANSCompel database, each entry corresponds to an individual *trans*/*cis* interaction within the context of a particular gene and thus contains information about two binding sites, two corresponding transcription factors and experiments confirming cooperative action between transcription factors.

ABS is a public database of known *cis*-regulatory binding sites identified in promoters of orthologous vertebrate genes that have been manually collated from the scientific literature (Blanco et al., 2006). In this database some 650 experimental binding sites from 68 transcription factors and 100 orthologous target genes in human, mouse, rat or chicken genome sequences have been documented. This tool allows computational predictions and promoter alignment information for each entry and is accessed through a simple and easy-to-use web interface; facilitating data retrieval and allowing different views of the information. One of the key features of this software is the inclusion of a customizable generator of artificial datasets based on the known sites contained in the whole collection and an evaluation tool to aid during the training and the assessment of various motif-finding programs.

JASPAR is an open-access database of annotated, high-quality, matrix-based TFBS profiles for multi-cellular eukaryotic organisms (Sandelin et al., 2004). The profiles were derived exclusively from sets of nucleotide sequences that were experimentally demonstrated to bind transcription factors. The database is accessible via a web-interface for browsing, searching and subset selection. The interface also includes an online sequence analysis utility and a suite of tools for genome-wide and comparative genome analysis of regulatory regions.

HTPSELEX is a public database providing access to primary and derived data from high-throughput SELEX experiments that were specifically designed in order to characterize the binding specificity of transcription factors (Jagannathan et al., 2006). The resource is primarily intended to serve computational biologists interested in building models of TFBSs from large sets of *cis*-regulatory sequences. For each experiment detailed in the database accurate information is provided about the protein material used, details of the wet lab protocol, an archive of sequencing trace files, assembled clone sequences and complete sets of *in vitro* selected protein-binding tags.

TRED is a database that stores both *cis*- and *trans*-regulatory elements and was designed to facilitate easy data access and to allow for the analysis of single-gene-based and genome-scale studies (Zhao et al., 2005). Distinguishing features of *TRED* include: relatively complete genome-wide promoter annotation for human, mouse and rat; availability of gene transcriptional regulation information including TFBSs and experimental evidence; data accuracy is ensured by hand curation; efficient user interface for easy and flexible data retrieval; and implementation of on-the-fly sequence analysis tools. *TRED* can provide good training datasets for further genome-wide *cis*-regulatory element prediction and annotation; assist detailed functional studies and facilitate the deciphering of gene regulatory networks.

Databases of known TFBSs can be used to detect the presence of protein-recognition elements in a given promoter, but only when the binding site of the relevant DNA-binding protein and its tolerance to mismatches *in vivo* is already known. Because this knowledge is currently limited to a small subset of transcription factors, much effort has been devoted to the discovery of regulatory motifs by comparative analysis of the DNA sequences of promoters. By finding conserved regions between multiple promoters, motifs can be identified with no prior knowledge of TFBSs. A number of models have emerged that achieve this by statistical overrepresentation. These algorithms function by aligning multiple untranslated regions from the entire genome and identifying sequences that are statistically significantly overrepresented in comparison to what it expected by random.

YMF is a program developed to identify novel TFBSs (not necessarily associated with a specific factor) in yeast by searching for statistically overrepresented motifs (Sinha et al., 2003; Sinha & Tompa, 2002). More specifically, *YMF* enumerates all motifs in the search space and is guaranteed to produce those motifs with the greatest z-scores.

SCORE is a computational method for identifying transcriptional *cis*-regulatory modules based on the observation that they often contain, in statistically improbable concentrations, multiple binding sites for the same transcription factor (Rebeiz et al., 2002). Using this method the authors conducted a genome-wide inventory of predicted binding sites for the Notch-regulated transcription factor Suppressor of Hairless, Su(H), in drosophila and found that the fly genome contains highly non-random clusters of Su(H) sites over a broad range of sequence intervals. They found that the most statistically significant clusters were very heavily enriched in both known and logical targets of Su(H) binding and regulation. The utility of the *SCORE* approach was validated by *in vivo* experiments showing that proper expression of the novel gene *Him* in adult muscle precursor cells depends both on Su(H) gene activity and sequences that include a previously unstudied cluster of four Su(H) sites, indicating that *Him* is a likely direct target of Su(H).

At present these tools are mainly applied in the study of lower eukaryotes where the genome is less complex and regulatory elements are easier to identify, extending these algorithms to the human genome has proven somewhat more difficult. In order to redress this issue a number of groups have shown that it is possible to mine the genome of higher eukaryotes by searching for conserved regulatory elements adjacent to transcription start site motifs such as TATA and CAAT boxes, e.g. as catalogued in the *DBTSS* resource (Suzuki et al. 2004; Suzuki et al., 2002), or one can search for putative *cis*-elements in CpG rich regions that are present in higher proportions in promoter sequences (Davuluri et al., 2001). Alternatively, with the co-emergence of microarray technology and the complete sequence of the human genome, it is now possible to search for potential TFBSs by comparing the upstream non-coding regions of multiple genes that show similar expression profiles under certain conditions. Gene sets for comparative analysis can be chosen based on clustering, e.g. hierarchical and k-means (Roth et al., 1998), from simple expression ratio (Bussemaker et al., 2001) or functional analysis of gene products (Jensen et al., 2000). This provides scientists with the opportunity to identify promoter elements that are responsive to certain environmental conditions, or those that play a key role in mediating the differentiation of certain tissues or those that may be particularly active in mediating pathologic phenotypes.

Phylogenetic footprinting, or comparative genomics, is now being applied to identify novel promoter elements by comparing the evolutionary conserved untranslated elements proximal to known genes from a variety of organisms. The availability of genome sequences between species has notably advanced comparative genomics and the understanding of evolutionary biology in general. The neutral theory of molecular evolution provides a framework for the identification of DNA sequences in genomes of different species. Its central hypothesis is that the vast majority of mutations in the genome are neutral with respect to the fitness of an organism. Whilst deleterious mutations are rapidly removed by selection, neutral mutations persist and follow a stochastic process of genetic drift through a population. Therefore, non-neutral DNA sequences (functional DNA sequences) must be conserved during evolution, whereas neutral mutations accumulate. Initial studies sufficiently demonstrated that the human genome could be adequately compared to the genomes of other organisms allowing for the efficient identification of homologous regions in functional DNA sequences.

Subsequently, a number of bioinformatics tools have emerged that operate by comparing non-coding regulatory sequences between the genomes of various organisms to enable the identification of conserved TFBSs that are significantly enriched in promoters of candidate genes or from clusters identified by microarray analysis; examples of these software suites are discussed below. Typically these tools work by aligning the upstream sequences of target genes between species thus identifying conserved regions that could potentially function as *cis*-regulatory elements and have consequently been applied in the elucidation of transcription regulatory networks in a variety of models.

TRAFAC is a Web-based application for analysis and display of a pair of DNA sequences with an emphasis on the detection of conserved TFBSs (Jegga et al., 2002). A number of programs are used to analyze the sequences and identify various genomic features (for example, exons, repeats, conserved regions, TFBSs). Repeat elements are masked out using *RepeatMasker* and the sequences are aligned using the *PipMaker-BLASTZ* algorithm. *MatInspector Professional* or *Match* (BioBase) is run to scan the sequences for TFBSs. *TRAFAC* then integrates analysis results from these applications and generates graphical outputs; termed the *Regulogram* and *Trafacgram.*

CORG comprises a catalogue of conserved non-coding sequence blocks that were initially computed based on statistically significant local suboptimal alignments of 15kb regions upstream of the translation start sites of some 10793 pairs of orthologous genes (Dieterich et al., 2003). The resulting conserved non-coding blocks were annotated with EST matches for easier detection of non-coding mRNA and with hits to known TFBSs. CORG data are accessible from the ENSEMBL web site via a DAS service as well as a specially developed web service for query and interactive visualization of the conserved blocks and their annotation.

CONSITE is a flexible suite of methods for the identification and visualization of conserved TFBSs (Lenhard et al., 2003). The system reports those putative TFBSs that are both situated in conserved regions and located as pairs of sites in equivalent positions in alignments between two orthologous sequences. An underlying collection of metazoan transcription-factor-binding profiles was assembled to facilitate the study. This approach results in a significant improvement in the detection of TFBSs because of an increased signal-to-noise ratio, as as the authors demonstrated with two sets of promoter sequences.

CONFAC enables the high-throughput identification of conserved TFBSs in the regulatory regions of hundreds of genes at a time (Karanam et al., (2004). The *CONFAC* software compares non-coding regulatory sequences between human and mouse genomes to enable identification of conserved TFBSs that are significantly enriched in promoters of gene clusters from microarray analyses compared to sets of unchanging control genes using a Mann–Whitney statistical test. The authors analysed random gene sets and demonstrated that using this approach, over 98% of TFBSs had false positive rates below 5%. As a proof-of-principle, the *CONFAC* software was validated using gene sets from four separate microarray studies and TFBSs were identified that are known to be functionally important for regulation of each of the four gene sets.

VAMP is a graphical user interface for both visualization and primary level analysis of molecular profiles obtained from functional genomics experimentation (La Rosa et al., 2006). It can be applied to datasets generated from Comparative Genomic Hybridisation (CGH) arrays, transcriptome arrays, Single Nucleotide Polymorphism arrays, loss of heterozygosity analysis (LOH), and Chromatin Immunoprecipitation experiments (ChIP-on-chip). The interface allows users to collate the results from these different types of studies and to view it in a convenient way. Several views are available, such as the classical CGH karyotype view or genome-wide multi-tumour comparisons. Many of the functionalities required for the analysis of CGH data are provided by the interface; including searches for recurrent regions of alterations, comparison to transcriptome data, correlation to clinical information, and the ability to view gene clusters in the context of genome structure.

CisMols Analyser allows for the filtering of candidate *cis*-element clusters based on phylogenetic conservation across multiple gene sets (Jegga et al., 2005). It was previously possible to achieve this for individual orthologue gene pairs, but combining data from *cis*-conservation and coordinate expression across multiple genes proved a more difficult task. To address this issue, the authors extended an orthologue gene pair database with additional analytical architecture to allow for the analysis and identification of maximal numbers of compositionally similar and phylogenetically conserved *cis*-regulatory element clusters from a list of user-selected genes. The system has been successfully tested with a series of functionally related and microarray profile-based co-expressed orthologue pairs of promoters and genes using known regulatory regions as training sets and co-expressed genes in the olfactory and immunohematologic systems as test sets. A significant amount of effort has been dedicated to the cataloguing of transcription factors and their corresponding *cis*-elements. More recently, these databases have been compiled with the aim to utilise them to unravel regulatory networks active in response to diverse stimuli.

PreMod was developed in an effort to identify *cis*-regulatory modules (CRM) active under specific environmental conditions (Blanchette et al., 2006; Ferretti et al., 2007). Starting from a set of predicted binding sites for more than 200 transcription factor families documented in the Transfac database (described above), the authors describe an algorithm relying on the principle that *cis*-regulatory modules (CRMs) generally contain several phylogenetically conserved binding sites for a small variety of transcription factors. The method allowed the prediction of more than 118,000 CRMs within the human genome. During this analysis, it was revealed that CRM density varies widely across the genome, with CRM-rich regions often being located near genes encoding transcription factors involved in development. Interestingly, in addition to showing enrichment near the 3′ end of genes, predicted CRMs were present in other regions more distant from genes. In this database, the tendency for certain transcription factors to bind modules located in specific regions was documented with respect to their target genes, and a number of transcription factors likely to be involved in tissue-specific regulation were identified.

CisView was developed to facilitate the analysis of gene regulatory regions of the mouse genome (Sharov et al., 2006). Its user interface is a browser and database of genome-wide potential TFBSs that were identified using 134 position-weight matrices and 219 sequence patterns from various sources. The output is presented with information about sequence conservation, neighbouring genes and their structures, GO annotations, protein domains, DNA repeats and CpG islands. The authors used this tool to analyse the distribution of TFBSs and revealed that many TFBSs were over-represented near transcription start sites. In the initial paper presenting the tool they also identified potential *cis*-regulatory modules defined as clusters of conserved TFBSs in the entire mouse genome. Out of 739,074 CRMs identified, 157,442 had a significantly higher regulatory potential score than semi-random sequences. The *CisView* browser provides a user-friendly computer environment for studying transcription regulation on a whole-genome scale and can also be used for interpreting microarray experiments and identifying putative targets of transcription factors.

BEARR is web browser software designed to assist biologists in efficiently carrying out the analysis of microarray data from studies of specific transcription factors (Vega et al., 2004). Batch Extraction and Analysis of *cis*-Regulatory Regions, or *BEARR*, accepts gene identifier lists from microarray data analysis tools and facilitates identification, extraction and analysis of regulatory regions from the large amount of data that is typically generated in these types of studies.

VISTA is a family of computational tools that was built to assist in the comparative analysis of DNA sequences (Dubchak & Ryaboy, 2006). These tools allow for the alignment of DNA sequences to facilitate the visualization of conservation levels and thus allow for the identification of highly conserved regions between species. Specifically, sequences can be analysed by browsing through pre-computed whole-genome alignments of vertebrates and other groups of organisms. Submission of sequences to *Genome VISTA* enables the user to align them to other whole genomes; whereas submission of two or more sequences to *mVISTA* allows for direct alignment. Submission of sequences to *Regulatory VISTA* is also possible and enables the predication of potential TFBSs (based on conservation within sequence alignments). All *VISTA* tools use standard algorithms for visualization and conservation analysis to make comparison of results from different programs more straightforward.

PromAn is a modular web-based tool dedicated to promoter analysis that integrates a number of different complementary databases, methods and programs (Lardenois et al., 2006). *PromAn* provides automatic analysis of a genomic region with minimal prior knowledge of the genomic sequence. Prediction programs and experimental databases are combined to locate the transcription start site (TSS) and the promoter region within a large genomic input sequence. TFBSs can be predicted using several public databases and user-defined motifs. Also, a phylogenetic footprinting strategy, combining multiple alignments of large genomic sequences and assignment of various scores reflecting the evolutionary selection pressure, allows for evaluation and ranking of TFBS predictions. *PromAn* results can be displayed in an interactive graphical user interface. It integrates all of this information to highlight active promoter regions, to identify among the huge number of TFBS predictions those which are the most likely to be potentially functional and to facilitate user refined analysis. Such an integrative approach is essential in the face of a growing number of tools dedicated to promoter analysis in order to propose hypotheses to direct further experimental validations.

CRSD is a comprehensive web server that can be applied in investigating complex regulatory behaviours involving gene expression signatures, microRNA regulatory signatures and transcription factor regulatory signatures (Liu et al., 2006). Six well-known and large-scale databases, including the human *UniGene*, mature microRNAs, putative promoter, *TRANSFAC*, *pathway* and *Gene Ontology* (GO) databases, were integrated to provide the comprehensive analysis in *CRSD*. Two new genome-wide databases, of microRNA and transcription factor regulatory signatures were also constructed and further integrated into *CRSD*. To accomplish the microarray data analysis at one go, several methods, including microarray data pre-treatment, statistical and clustering analysis, iterative enrichment analysis and motif discovery, were closely integrated in the web server.

MPromDb is a database that integrates gene promoters with experimentally supported annotation of transcription start sites, *cis*-regulatory elements, CpG islands and chromatin immunoprecipitation microarray (ChIP-chip) experimental results within an intuitively designed interface (Sun et al., 2006). Its initial release contained information on 36,407 promoters and first exons, 3,739 TFBSs and 224 transcription factors; with links to *PubMed* and *GenBank* references. Target promoters of transcription factors that have been identified by ChIP-chip assay are also integrated into the database and thus serving as a portal for genome-wide promoter analysis of data generated by ChIP-chip experimental studies.

A comprehensive list of the all the databases described above with a summary of their features and a reference to the original citation are shown in table 1.

Each of the aforementioned databases can be used when searching for potential regulatory sequences for inclusion in the design of synthetic promoters. Indeed, these resources can be used in order to identify *cis*-regulatory elements that may play a role in the formation of a particular cellular phenotype, or those that may be important in driving differentiation in developing organs. Synpromics, an emerging synthetic biology company recently incorporated in the United Kingdom, has cleverly utilised these tools in developing a proprietary method of synthetic promoter production where identified elements are incorporated into the design of promoters that are able to specifically regulate gene expression in a particular cellular phenotype. This method harnesses a cell's gene expression profile in order to facilitate the design of highly specific and efficient promoters. The result is a range of promoters that are inducible, tissue (or cell)-specific, active in response to a particular pathogen, chemical or biological agent and even able to mediate gene expression only under certain pathological conditions, such as cancer. Indeed, Synpromics has successfully generated a range of synthetic promoters that specifically drive high levels of gene expression in colorectal cancer and are looking to apply these promoters in the development of safer gene therapies (*manuscript in preparation*).

Resource	Description	Citation
DBTSS	Database of transcriptional start sites	Suzuki et al., (2002)
TRAFAC	Conserved *cis*-element search tool	Jegga et al., (2002)
TRANSCompel	Database of composite regulatory elements	Kel-Margoulis et al., (2002)
TRANSFAC	Eukaryotic transcription factor database	Matys et al., (2003)
Phylofoot	Tools for phylogenetic footprinting purposes	Lenhard et al., (2003)
CORG	Multi-species DNA comparison and annotation	Dieterich et al., (2003)
CONSITE	Explores *trans*-factor binding sites from two species	Lenhard et al., (2003)
CONFAC	Conserved TFBS finder	Karanam et al., (2004)
CisMols	Identifies *cis*-regulatory modules from inputed data	Jegga et al., (2005)
TRED	Catalogue of transcription regulatory elements	Zhao et al., (2005)
Oncomine	Repository and analysis of cancer microarray data	Rhodes et al., (2005)
ABS	Database of regulatory elements	Blanco et al., (2006)
JASPAR	Database of regulatory elements	Sandelin et al., (2004)
HTPSELEX	Database of composite regulatory elements	Jagannathan et al., (2006)
PReMod	Database of transcriptional regulatory modules in the human genome	Blanchette et al., (2006)
CisView	Browser of regulatory motifs and regions in the genome	Sharov et al., (2006)
BEARR	Batch extraction algorithm for microarray data analysis	Vega et al., 2004)
VISTA	Align and compare sequences from multiple species	Dubchak et al., (2006)
PromAn	Promoter analysis by integrating a variety of databases	Lardenois et al., (2006)

Table 1. Bioinformatics tools used for the identification of *cis*-regulatory elements

Importantly, synthetic promoters often mediate a level of gene expression with much greater efficiency than that seen with viral promoters, such as CMV, or compared to naturally occurring promoters within the genome. Given that the entire Biotech industry is centred on the regulation of gene expression, it is likely that synthetic promoters will eventually replace all naturally-occurring sequences in use today and help drive the growth of the synthetic biology sector in the coming decades.

6. Conclusion

In summary, synthetic promoters have emerged over the past two decades as excellent tools facilitating the identification of important structural features in naturally occurring promoter sequences and allowing enhanced and more restrictive regulation of gene expression. A number of early studies revealed that it was possible to combine the *cis*-regulatory elements from promoters of a few tissue-specific genes and use these as building blocks to generate shorter, more efficient tissue-specific promoters. Several simple methodologies to achieve this emerged and have been applied in a multitude of organisms; including plant, bacteria, yeast, viral and mammalian systems.

Recent advances in bioinformatics and the emergence of a plethora of tools specifically designed at unravelling transcription programs has also facilitated the design of highly-specific synthetic promoters that can drive efficient gene expression in a tightly regulated manner. Changes in a cell's gene expression profile can be monitored and the transcription programs underpinning that change delineated and the corresponding *cis*-regulatory modules can be used to construct synthetic promoters whose activity is restricted to individual cell types, or to single cells subject to particular environmental conditions. This has allowed researchers to design promoters that are active in diseased cells or in tissues treated with a particular biological or chemical agent; or active in cells infected with distinct pathogens.

A number of institutions, such as Synpromics, have taken advantage of these advances and are now working to apply synthetic promoter technology to the enhanced production of biologics for use in biopharmaceutical, greentech and agricultural applications; the development of new gene therapies; and in the design of a novel class of molecular diagnostics. As the synthetic biology field continues to develop into a multi-billion dollar industry, synthetic promoter technology is likely to remain at the heart of this ever-expanding and exciting arena.

7. References

Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganière, J., Lefèbvre, C., et al. (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Research*, Vol. 16, No. 5, May 2006, pp. 656-668. ISSN 1088-9051

Blanco, E., Farré, D., Albà, M. M., Messeguer, X., & Guigó, R. (2006). ABS: a database of Annotated regulatory Binding Sites from orthologous promoters. *Nucleic Acids Research*, Vol. 34, January 2006, pp. D63-67. ISSN 0305-1048

Bussemaker, H. J., Li, H., & Siggia, E D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, Vol. 27, No. 2, February 2001, pp. 167-71. ISSN 1061-4036

Chow, Y. H., O'Brodovich, H., Plumb, J., Wen, Y., Sohn, K. J., Lu, Z., et al. (1997). Development of an epithelium-specific expression cassette with human DNA regulatory elements for transgene expression in lung airways. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 94, No. 26, December 1997, pp. 14695-14700. ISSN 1091-6490

Dai, C., McAninch, R. E., & Sutton, R. E. (2004). Identification of synthetic endothelial cell-specific promoters by use of a high-throughput screen. *Journal of Virology*, Vol. 78, No. 12, June 2004, pp. 6209-6221. ISSN 0022-538X

Davuluri, R V, Grosse, I., & Zhang, M Q. (2001). Computational identification of promoters and first exons in the human genome. *Nature Genetics*, Vol. 29, No. 4, December 2001, pp. 412-417. ISSN 1061-4036

Dieterich, C., Wang, H., Rateitschak, K., Luz, H., & Vingron, M. (2003). CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Research*, Vol. 31, No. 1, January 2003, pp. 55-57. ISSN 0305-1048

Dubchak, I., & Ryaboy, D. V. (2006). VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. *Methods in Molecular Biology*, Vol. 338, April 2006, pp. 69-89. ISSN 1064-3745

Edelman, G. M., Meech, R., Owens, G. C., & Jones, F. S. (2000). Synthetic promoter elements obtained by nucleotide sequence variation and selection for activity. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 97, No. 7, March 2000, pp. 3038-3043. ISSN 1091-6490

Ferretti, V., Poitras, C., Bergeron, D., Coulombe, B., Robert, F., & Blanchette, M. (2007). PReMod: a database of genome-wide mammalian *cis*-regulatory module predictions. *Nucleic Acids Research*, Vol. 35, January 2007, pp. D122-126. ISSN 0305-1048

Gertz, J., & Cohen, B. A. (2009). Environment-specific combinatorial *cis*-regulation in synthetic promoters. *Molecular Systems Biology*, Vol. 5, February 2009, pp. 244. ISSN 1744-4292

Gertz, J., Siggia, Eric D, & Cohen, B. A. (2009). Analysis of combinatorial *cis*-regulation in synthetic and genomic promoters. *Nature*, Vol. 457, No. 7226, July 2009, pp. 215-218. ISSN: 0028-0836

Hammer, K., Mijakovic, I., & Jensen, P. R. (2006). Synthetic promoter libraries--tuning of gene expression. *Trends in Biotechnology*, Vol. 24, No. 2, February 2006, pp. 53-55. ISSN 0167-7799

Han, J., McLane, B., Kim, E.-H., Yoon, J.-W., & Jun, H.-S. (2010). Remission of Diabetes by Insulin Gene Therapy Using a Hepatocyte-specific and Glucose-responsive Synthetic Promoter. *Molecular Therapy*, Vol. 19, No. 3, March 2010, pp.470-478. ISSN 1525-0016

He, W., Qiang, M., Ma, W., Valente, A. J., Quinones, M. P., Wang, W., et al. (2006). Development of a synthetic promoter for macrophage gene therapy. *Human Gene Therapy*, Vol. 17, No. 9, September 2006, pp. 949-959. ISSN 1043-0342

Hunziker, A., Tuboly, C., Horváth, P., Krishna, S., & Semsey, S. (2010). Genetic flexibility of regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 107, No. 29, July 2010, pp. 12998-13003. ISSN 1091-6490

Hwang, D. Y., Carlezon, W. A., Isacson, O., & Kim, K. S. (2001). A high-efficiency synthetic promoter that drives transgene expression selectively in noradrenergic neurons. *Human Gene Therapy*, Vol. 12, No. 14, September 2001, pp.1731-1740. ISSN 1043-0342

Jacquet, M. A., Ehrlich, R., & Reiss, C. (1989). In vivo gene expression directed by synthetic promoter constructions restricted to the -10 and -35 consensus hexamers of E. coli. *Nucleic Acids Research*, Vol. 17, No. 8, April 1989, pp. 2933-2945. ISSN 0305-1048

Jagannathan, V., Roulet, E., Delorenzi, M., & Bucher, P. (2006). HTPSELEX--a database of high-throughput SELEX libraries for TFBSs. *Nucleic Acids Research*, Vol. 34, January 2006, pp. D90-94. ISSN 0305-1048

Jegga, A. G., Sherwood, S. P., Carman, J. W., Pinski, A. T., Phillips, J. L., Pestian, J. P., et al. (2002). Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Research*, Vol. 12, No. 9, September 2002, pp. 1408-1417. ISSN 1088-9051

Jegga, A. G., Gupta, A., Gowrisankar, S., Deshmukh, M. A., Connolly, S., Finley, K., et al. (2005). CisMols Analyzer: identification of compositionally similar *cis*-element clusters in ortholog conserved regions of coordinately expressed genes. *Nucleic Acids Research*, Vol. 33, July 2005, pp. W408-411. ISSN 0305-1048

Jensen, L. J., & Knudsen, S. (2000). Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, Vol. 16, No. 4, April 2000, pp. 326-333. ISSN 1460-2059

Kamine, J., Subramanian, T., & Chinnadurai, G. (1991). Sp1-dependent activation of a synthetic promoter by human immunodeficiency virus type 1 Tat protein. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 88, No. 19, October 1991, pp. 8510-8514. ISSN 1091-6490

Karanam, S., & Moreno, C. S. (2004). CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets. *Nucleic Acids Research*, Vol. 32, July 2004, pp. W475-484. ISSN 0305-1048

Kel-Margoulis, Olga V, Kel, Alexander E, Reuter, Ingmar, Deineko, I. V., & Wingender, Edgar. (2002). TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Research*, Vol. 30, No. 1, January 2002, pp. 332-334. ISSN 0305-1048

Kinkhabwala, A., & Guet, C. C. (2008). Uncovering *cis* regulatory codes using synthetic promoter shuffling. *PloS One*, Vol. 3, No. 4, April 2008, pp. e2030. ISSN 1932-6203

La Rosa, P., Viara, E., Hupé, P., Pierron, G., Liva, S., Neuvial, P., et al. (2006). VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles. *Bioinformatics*, Vol. 22, No. 17, September 2006, pp. 2066-2073. ISSN 1460-2059

Lardenois, A., Chalmel, F., Bianchetti, L., Sahel, J.-A., Léveillard, T., & Poch, O. (2006). PromAn: an integrated knowledge-based web server dedicated to promoter analysis. *Nucleic Acids Research*, Vol. 34, July 2006, pp. W578-83. ISSN 0305-1048

Lemken, M.-L., Wybranietz, W.-A., Schmidt, U., Graepler, F., Armeanu, S., Bitzer, M., et al. (2005). Expression liver-directed genes by employing synthetic transcriptional control units. *World Journal of Gastroenterology*, Vol. 11, No. 34, September 2005, pp. 5295-5302. ISSN 1007-9327

Lenhard, B., Sandelin, A., Mendoza, L., Engström, P., Jareborg, N., & Wasserman, W. W. (2003). Identification of conserved regulatory elements by comparative genome analysis. *Journal of Biology*, Vol. 2, No. 2, May 2003, pp. 13. ISSN 1475-4924

Li, X., Eastman, E. M., Schwartz, R. J., & Draghia-Akli, R. (1999). Synthetic muscle promoters: activities exceeding naturally occurring regulatory sequences. *Nature Biotechnology*, Vol. 17, No. 3, January 1999, pp. 241-245. ISSN: 1087-0156

Liu, C.-C., Lin, C.-C., Chen, W.-S. E., Chen, H.-Y., Chang, P.-C., Chen, J. J. W., et al. (2006). CRSD: a comprehensive web server for composite regulatory signature discovery. *Nucleic Acids Research*, Vol. 34, July 2006, pp. W571-7. ISSN 0305-1048

Liu, M.-L., Oh, J. S., An, S. S., Pennant, W. A., Kim, H. J., Gwak, S.-J., et al. (2010). Controlled nonviral gene delivery and expression using stable neural stem cell line transfected with a hypoxia-inducible gene expression system. *The Journal of Gene Medicine*, Vol. 12, No. 12, December 2010, pp. 990-1001. ISSN 1521-2254

Logg, C. R., Logg, A., Matusik, R. J., Bochner, B. H., & Kasahara, N. (2002). Tissue-specific transcriptional targeting of a replication-competent retroviral vector. *Journal of Virology*, Vol. 76, No. 24, December 2002, pp. 12783-12791. ISSN 0022-538X

Mader, S., & White, J. H. (1993). A steroid-inducible promoter for the controlled overexpression of cloned genes in eukaryotic cells. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 90, No. 12, June 1993, pp. 5603-5607. ISSN 1091-6490

Masson, J. M., & Miller, J. H. (1986). Expression of synthetic suppressor tRNA genes under the control of a synthetic promoter. *Gene*, Vol. 47, No. 2-3, February 1986, pp.179-83. ISSN 0378-1119

Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., et al. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, Vol. 31, No. 1, January 2003, pp. 374-378. ISSN 0305-1048

Mogno, I., Vallania, F., Mitra, R. D., & Cohen, B. A. (2010). TATA is a modular component of synthetic promoters. *Genome Research*, Vol. 20, No. 10, October 2010, pp. 1391-1397. ISSN 1088-9051

Ramon, A., & Smith, H. O. (2010). Single-step linker-based combinatorial assembly of promoter and gene cassettes for pathway engineering. *Biotechnology Letters*, Vol 33, No. 3, March 2010, pp. 549-555. ISSN 0141-5492

Rebeiz, M., Reeves, N. L., & Posakony, J. W. (2002). SCORE: a computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 99, No. 15, July 2002, pp. 9888-9893. ISSN 1091-6490

Roberts, M. L. (2007). A method for the construction of cancer-specific promoters using functional genomics. WIPO WO/2008/107725.

Roberts, M. L., & Kottaridis S. D. (2007). Interpreting microarray data: towards the complete bioinformatics toolkit for cancer. Cancer Genomics and Proteomics, Vol 4, No. 4, July-August 2007, pp 301-308. ISSN: 1109-6535.

Roth, F. P., Hughes, J. D., Estep, P. W., & Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, Vol. 16, No. 10, October 1998, pp. 939-945. ISSN: 1087-0156

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., & Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, Vol. 32, January 2004, pp. D91-94. ISSN 0305-1048

Santilli, G., Almarza, E., Brendel, C., Choi, U., Beilin, C., Blundell, M. P., et al. (2011). Biochemical correction of X-CGD by a novel chimeric promoter regulating high levels of transgene expression in myeloid cells. *Molecular Therapy*, Vol. 19, No. 1, January 2011, pp. 122-132. ISSN 1525-0016

Segal, E., Friedman, N., Koller, D., & Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. *Nature Genetics*, Vol. 36, No. 10, October 2004, pp. 1090-1098. ISSN 1061-4036

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., et al. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, Vol. 34, No. 2, June 2003, pp. 166-176. ISSN 1061-4036

Sharov, A. A., Dudekula, D. B., & Ko, M. S. H. (2006). CisView: a browser and database of *cis*-regulatory modules predicted in the mouse genome. *DNA Research*, Vol. 13, No. 3, June 2006, pp. 123-134. ISSN 1340-2838

Sinha, S., & Tompa, M. (2002). Discovery of novel TFBSs by statistical overrepresentation. *Nucleic Acids Research*, Vol. 30, No. 24, December 2002, pp. 5549-5560. ISSN 0305-1048

Sinha, S., & Tompa, M. (2003). YMF: A program for discovery of novel TFBSs by statistical overrepresentation. *Nucleic Acids Research*, Vol. 31, No. 13, July 2003, pp. 3586-3588. ISSN 0305-1048

Sleight, S. C., Bartley, B. A., Lieviant, J. A., & Sauro, H. M. (2010). Designing and engineering evolutionary robust genetic circuits. *Journal of Biological Engineering*, Vol. 4, No. 1, November 2010, pp. 12. ISSN 1754-1611

Sun, H., Palaniswamy, S. K., Pohar, T. T., Jin, V. X., Huang, T. H.-M., & Davuluri, Ramana V. (2006). MPromDb: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-chip experimental data. *Nucleic Acids Research*, Vol. 34, January 2006, pp. D98-103. ISSN 0305-1048

Suzuki, Y., Yamashita, R., Nakai, K., & Sugano, S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Research*, Vol. 30, No. 1, January 2002, pp. 328-331. ISSN 0305-1048

Suzuki, Y., Yamashita, R., Sugano, S., & Nakai, K. (2004). DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Research*, Vol. 32, January 2004, pp. D78-81. ISSN 0305-1048

Szymanski, P., Anwer, K., & Sullivan, S. M. (2006). Development and characterization of a synthetic promoter for selective expression in proliferating endothelial cells. *The Journal of Gene Medicine*, Vol. 8, No. 4, April 2006, pp. 514-523. ISSN 1521-2254

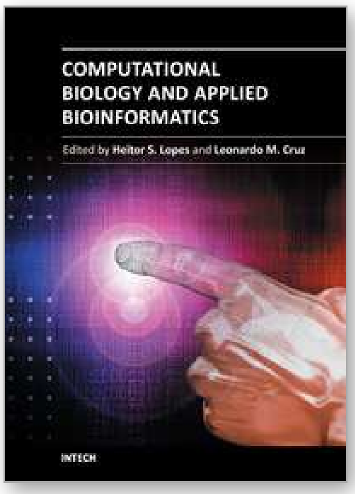
Trumble, W. R., Sherf, B. A., Reasoner, J. L., Seward, P. D., Denovan, B. A., Douthart, R. J., et al. (1992). Protein expression from an Escherichia coli/Bacillus subtilis multifunctional shuttle plasmid with synthetic promoter sequences. *Protein Expression and Purification*, Vol. 3, No. 3, June 1992, pp. 169-177. ISSN 1046-5928

Vega, V. B., Bangarusamy, D. K., Miller, L. D., Liu, E. T., & Lin, C.-Y. (2004). BEARR: Batch Extraction and Analysis of *cis*-Regulatory Regions. *Nucleic Acids Research*, Vol 32, July 2004, pp. W257-260. ISSN 0305-1048

Venter, M. (2007). Synthetic promoters: genetic control through *cis* engineering. *Trends in Plant Science*, Vol. 12, No. 3, March 2007, pp. 118-124. ISSN 1360-1385

Wingender, E. (1988). Compilation of transcription regulating proteins. *Nucleic Acids Research*, Vol. 16, No. 5, March 1988, pp. 1879-1902. ISSN 0305-1048

Zhao, F., Xuan, Z., Liu, L., & Zhang, Michael Q. (2005). TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. *Nucleic Acids Research*, Vol. 33, January 2005, pp. D103-107. ISSN 0305-1048


COMPUTATIONAL BIOLOGY AND APPLIED BIOINFORMATICS
Edited by Heitor S. Lopes and Leonardo M. Cruz
INTECH

Computational Biology and Applied Bioinformatics
Edited by Prof. Heitor Lopes

ISBN 978-953-307-629-4
Hard cover, 442 pages
Publisher InTech
Published online 02, September, 2011
Published in print edition September, 2011

Nowadays it is difficult to imagine an area of knowledge that can continue developing without the use of computers and informatics. It is not different with biology, that has seen an unpredictable growth in recent decades, with the rise of a new discipline, bioinformatics, bringing together molecular biology, biotechnology and information technology. More recently, the development of high throughput techniques, such as microarray, mass spectrometry and DNA sequencing, has increased the need of computational support to collect, store, retrieve, analyze, and correlate huge data sets of complex information. On the other hand, the growth of the computational power for processing and storage has also increased the necessity for deeper knowledge in the field. The development of bioinformatics has allowed now the emergence of systems biology, the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of a living being. This book presents some theoretical issues, reviews, and a variety of bioinformatics applications. For better understanding, the chapters were grouped in two parts. In Part I, the chapters are more oriented towards literature review and theoretical issues. Part II consists of application-oriented chapters that report case studies in which a specific biological problem is treated with bioinformatics tools.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Michael L. Roberts (2011). The Use of Functional Genomics in Synthetic Promoter Design, Computational Biology and Applied Bioinformatics, Prof. Heitor Lopes (Ed.), ISBN: 978-953-307-629-4, InTech, Available from: http://www.intechopen.com/books/computational-biology-and-applied-bioinformatics/the-use-of-functional-genomics-in-synthetic-promoter-design

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen