USING COMPUTATIONAL APPROACH IN UNDERSTANDING GENE REGULATORY NETWORKS FOR ANTIMICROBIAL PEPTIDE CODING GENES

MANISHA BRAHMACHARY
(M. *Sc., Indian Institute of Technology, Roorkee, India*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOCHEMISTRY

NATIONAL UNIVERSITY OF SINGAPORE

2006

ACKNOWLEDGEMENTS

Throughout my Ph.D. candidature, I have been supported by friends and family members to complete this thesis. So, it is with deep gratitude that I express my heartfelt appreciation to the following:

- Almighty God who stood by me always and held my hand in the face of adversity.
- Professor Vladimir Bajic, my supervisor and mentor, who guided me throughout this process and with whom numerous discussions on various scientific aspects of the project strengthened my analytical skill and expertise in sequence analysis.
- A/P Tan Tin Wee, my co-supervisor, who gave me advice and support which motivated me to pursue this Ph.D.
- Yang Liang, Huang Enli and Sin Lam, Vidhu and Krishnan for their computing assistance in my research.
- Asif, Paul, Rajesh, Dr. Bijaya for their critique and discussion of my work and companionship at I^2R.
- My father and mother for their care, support and going the extra mile to help me hold on in difficult times.
- My husband for his support and patience

My deepest and sincere gratitude,

Manisha Brahmachary

August, 2006

TABLE OF CONTENTS

SUMMARY

Antimicrobial peptides (AMPs) play a key role in the innate immune response. They can be ubiquitously found in a wide range of eukaryotes including mammals, amphibians, insects, plants, and protozoa. In lower organisms, AMPs function merely as antibiotics by permeabilizing cell membranes and lysing invading microbes. However, during evolution these peptides have become multifunctional molecules acting in the complex networks of higher organisms with additional properties such as having a mitogenic activity, antitumor activity or playing a role in adaptive immune responses. Hence, the AMPs are interesting targets to analyze transcriptional regulatory networks as their involvement in diverse pathways suggests. Understanding transcription regulation of any class of gene is a mammoth task, which can be approached from many angles. The author has focused on promoter region analysis of AMP genes, specifically to find transcription factor binding site motifs. The questions that were asked in the beginning of the thesis were, what are the promoter elements that regulate transcription of different AMP genes? Are they common across different AMP genes or specific to each AMP gene or AMP gene group? Are the promoter elements conserved across different species of an AMP gene group? Can promoter element modules be created out of these promoter elements? Can new AMP genes be found using the non-homology, promoter analysis based approach? This thesis has attempted to answer these questions by using examples of several AMP gene families. To be able to address the questions raised for this thesis, the author employed an array of computational biology techniques (sequence analysis based), supported by statistical evidence in a stepwise manner. The thesis begins with the creation of an antimicrobial peptide database (Chapter 3) that proved to be a good resource for the

research done for this thesis. Some prominent AMP families were analyzed in depth at peptide level and Hidden Markov Model (HMM) method was employed as a prediction tool to elucidate plausible important functional residues of some AMP families (Chapter 4). The author further delved into the gene level of AMPs and used the antimicrobial peptide database as a starting point to narrow down the families to work on for transcription regulation. The author has also collaborated with RIKEN Institute, Japan, for this research and used FANTOM full-length cDNA repository from RIKEN that was unpublished data resource at the time this research began.

Ab-initio motif finding method was used to find novel promoter elements (PEs*). The author was able to find common and different PEs between different species for AMP families (Chapter 5). The common, conserved PEs were used to develop specific models of promoters of co-regulated genes or genes having similar function (Chapter 6). These models were then used to search across the human promoter data for potentially new genes that have high possibility of being co-expressed as the target AMP gene group (Chapter 7). The search across the promoter regions of the human genome was done with the idea that the outcome will be a set of genes and/or new AMP genes themselves. Thus, this approach facilitates unfolding the relationship of AMP genes with other genes of the same pathway and helps us understand parts and functions of the underlying gene networks. This indirectly enriches the knowledge about the responses that cells generate while reacting to pathogen invasion and potentially can help in designing better antimicrobial drugs.

*PE is abbreviation for Promoter Element, which has been used interchangebly with TFBS in this thesis

LIST OF TABLES

Supplementary Tables

LIST OF FIGURES

List of Abbreviations

AMP:	Antimicrobial peptide
DEFA1:	Alpha defensin 1
DEFA3:	Alpha defensin 3
DMB:	Dragon Motif Builder
EM:	Expectation Maximization (algorithm)
EST:	Expressed Sequence Tag
FANTOM:	Functional Annotation of the mouse
FlcDNA:	Full length cDNA
GO:	Gene Ontology
GRN:	Gene Regulatory Network
HMM:	Hidden Markov Model
HNP-1:	Neutrophil defensin 1
HNP-3	Neutrophil defensin 3
NHR:	Nuclear Hormone Receptor
PE:	Promoter Element (used interchangeably as Transcription Factor Binding Sites (TFBS)
Penk1:	Preproenkephalin 1
PWM:	Position Weight Matrix
SAGE:	Serial Analysis of Gene Expression
TC:	Tag Cluster
TF:	Transcription Factor
TFBS:	Transcription Factor Binding Site

Part I Chapter 1: Introduction

The art of being wise is knowing what to overlook.

(William James)

1.1 Background on AMPs

Antimicrobial peptides (AMPs) are integral components of innate immunity in many organisms. They may be broadly classified into two classes, those that are directly anti-microbial, and those that are derived by proteolytic cleavage of a precursor. (Pazgier *et al.*, 2006, Li *et al.*, 2006, Shinnar *et al.*, 2003 , Ibrahim *et al.*, 2005 , von Horsten *et al.*, 2002).

Mammals produce many different antimicrobial peptides that are active against a broad spectrum of pathogens, including Gram-positive and Gram-negative bacteria, rickettsia, protozoans, fungi and some viruses (Hancock and Diamond, 2000)

Many AMPs are also involved in functions not directly associated with the innate immune response. For example, under normal physiological conditions, hepcidin is an important regulator of hepatic iron homeostasis, but at least in zebra fish it also acts as AMP (Shike *et al.*, 2004). Another AMP, the neutrophil granule derived peptide cap37, which binds to Gram-negative bacterial endotoxins, also acts as signaling molecule causing the up-regulation of protein kinase C activity (Kamysz *et al.*, 2003). Individual AMPs may have distinct functions in different locations (for example, at mucosal surfaces or in phagocytes), and must be regulated so as to be available when the pathogen challenge is presented. This instigates an interesting research problem, which is, to understand underlying transcriptional players for different families of AMP genes and networks in which they maybe involved and regulated.

1.2 Research issues investigated in this thesis

AMPs are of commercial and academic interest due to their unique sequence properties and ability to attack an array of pathogens. Realizing the importance of these groups of genes, gene discovery efforts have been undertaken by many groups. For example, efforts were directed to the computational discovery of beta defensin producing genes (Scheetz *et al.*, 2002, Schutte *et al.*, 2002). The method used is based on a similarity approach associated with HMM search and BLAST search of EST sequences mapped to confirm the transcription of these genes. However, this approach has some inherent limitations as both BLAST and HMMER analyses could not identify all known beta defensin genes, even not all used in the training of HMMER (Schutte *et al.*, 2002). This was due to the fact that AMPs are highly diverse peptide sequences even within the same family and species (Maxwell *et al.*, 2003, Tennessen, 2005). Hence, similarity can be very low in which case it is difficult to decide if putative hits obtained with low similarity can be considered being new AMPs.

The discovery of new AMP coding genes (AMPcgs) can be considered a special case of the general gene discovery problem. The existing experimental and computational methods (Xiang and Chen, 2000, Iida and Nishimura, 2002, Maggio and Ramnarayan, 2001, Zhang, 2002) are not specifically tuned to this gene class, which reduces chances for targeted search for AMP genes. For example, the common approach that can be used to search for new AMP members is homology search by tools like BLAST against known and 'artificial' (DNA translated) peptide sequences (Xiao *et al.*, 2004, Zaballos *et al.*, 2004). While this approach is widely used, it suffers a serious problem related to the level of similarity through which one can infer that the predicted peptide belongs to the target

group. A new methodology for computational gene discovery has been proposed and used recently for some specific classes of genes (Frech *et al.*, 1997, Wasserman and Fickett, 1998) based on the concept of modelling of the gene's promoter region. This approach seems reasonable to use for the purpose of AMP gene discovery as literature reviews suggest that the promoter regions of the highly diverse AMPs are fairly conserved (Ganz, 2003). This can suitably complement homology based gene identification. This approach also facilitates in unfolding of possible new association of genes with other genes (in terms of co-regulation) of the same pathway and unearthing parts and functions of the underlying gene networks which earlier have not been reported (Cohen *et al.*, 2006, Dohr *et al.*, 2005).

In this study, the major aim has been to use computational approaches to find the underlying PEs i.e. the transcription factor binding sites (TFBSs) and their organization across different AMP families. This is a challenging computational problem because of the difficulty finding true TFBSs in promoter regions .The TFBSs in promoter regions are very short motifs and their sequence variability has not been very well understood. Secondly, the promoter regions of genes can be several hundred to thousand base pairs long and the TFBSs can lie anywhere across the region. Finding true positive TFBSs has been the aim of many groups working on algorithms to predict the TFBS motifs (Hertz and Stormo, 1999, Frith *et al.*, 2004, Bailey and Elkan, 1995). The TFBS motifs, which are cis-elements and are present nearby each other in the promoter region, can be grouped into modules. Some of these modules* have been observed to be conserved across different classes of genes or across different species for the same genes. This phenomenon is particularly seen in genes of belonging to a particular classes and having

similar functions that co-express together under specific conditions (Werner *et al.*, 2003, Werner, 2003, Werner, 2002). Thus, genes under the same conditions have similar TFBS patterns contained in their promoter regions. These TFBS patterns can be used to develop specific models of promoters of co-regulated genes and these models can be used to search across genome for potential new genes that also have high chance of being co-expressed as the target gene group (Werner, 2001). Genes predicted on the basis of derived promoter models of the target AMP gene group are expected to be genes that could be part of the same pathway in which an AMP participates directly or indirectly (Niyonsaba *et al.*, 2003, Wang *et al.*, 2003, Moon *et al.*, 2002). and some could be AMP genes.

Using promoter region analysis to find new AMP genes and co-regulated genes is a first of its kind approach in the field of antimicrobial peptides. The results of this analysis can guide the way for experimental validation of the predicted set of genes. This thesis attempts to add knowledge to the understanding of transcriptional regulation of AMPs based on computational methods.

In order to achieve this primary objective, the secondary objectives of this thesis include (a) building a comprehensive repository of AMPs and (b) integrating analysis tool for sequence based classification. These objectives lay the foundations that would facilitate future wider systematic studies of the various AMP families in addition to the goals of this thesis in exploring the promoter elements of AMP.

1.3 Objectives of this thesis

Large-scale analysis of antimicrobial peptide genes at promoter level provides a global view on their transcriptional regulation level. This analysis in turn can support experimental studies by assisting in planning critical experiments and, when properly used, it can significantly improve the efficacy of experimental studies to understand transcriptional regulation. This research area is important for increasing our insight and knowledge about the little known area of transcriptional regulation of AMPs. In general, AMPs display an array of diverse functions and new information about their transcriptional regulation can help us understand their role and position in innate immunity, adaptive immunity and other related pathways in a better way. This would in turn have long-term implications in their role as potential drug candidates.

The first step towards executing a systematic data mining strategy to deduce novel insights into huge amount of biological data is to provide an adequate data management pipeline. Thus, consolidating the scattered data on antimicrobial peptides into a centralized database is a prerequisite for a systematic large-scale analysis. Information gained from such analysis is useful for developing new analytical tools for study of novel antimicrobial sequences.

Therefore, the specific objectives of this thesis were to:

1. Build a database of antimicrobial peptides with integrated query, extraction and sequence analysis tools, (Chapter 3, 4)
2. Extract and analyze the promoter dataset of AMP genes and find the key regulatory elements that are playing a role, (Chapter 5)
3. Develop promoter models of AMP genes for several AMP families, (Chapter 6) and

4. Use promoter models to search across human promoter data for (Chapter 7)
    a) detection of new co-regulated genes, and
    b) deciphering parts of gene networks of which AMP genes are members.

1.4 Contribution of this thesis

AMP-coding genes and their products have been extensively analyzed with regard to evolution (Crovella *et al.*, 2005 Patil *et al.*, 2004, Xiao *et al.*, 2004, Rodriguez de la Vega and Possani, 2005). Functional studies focusing on biochemical and immunological characterization have been performed on individual members (Krause *et al.*, 2003 Kragol *et al.*, 2001, Risso, 2000, Selsted *et al.*, 1993). However, until now there has not been any comprehensive characterization of promoter regions among all mammalian AMPs. This study is unique in scale and methodology. The author has employed a combination of computational methods and proper statistical testing and, 1) identified in promoter regions of 77 genes representing 22 AMP families known and novel transcription factor binding motifs, 2) their combinations and conserved modules, and 3) linked them according to biological functions in context of the AMPs.

The author's original contributions to the field of antimicrobial peptides include:

1) Organizing a large and unique data set of ~1788 entries of antimicrobial peptides from public databases and literature and creating a web-accessible, publicly available database (**http://research.i2r.a-star.edu.sg/Templar/DB/ANTIMIC)**. This database of antimicrobial peptides is the most comprehensive resource (eukaryotic and prokaryotic) for researchers to identify antimicrobial peptides and

analyze their sequence which otherwise would involve multiple querying of other databases. Integration of Hidden Markov Model (HMM) based tool and using it to find the potentially important residues of functional importance in certain AMP families.

2) Identifying common and specific putative regulatory elements (TFBS motifs) within the AMPcg's promoter regions. These findings have been supported by literature evidence wherever possible.
3) Developing promoter models of several AMP gene groups. To the best of the author's knowledge and based on the literature search, there have been no attempts to model promoters of AMPcgs.
4) Identifying likely co-regulated AMPcgs using AMP promoter models based on a scan across promoter regions of the human genome and determining parts of potential transcription regulatory networks in which some of the AMP genes are possibly involved.
5) Providing a functional analysis of the genes so identified and their relation to particular gene networks.

1.5 A summary of the thesis

This thesis consists of three parts. Part I provides an introduction to the thesis, in terms of the importance of antimicrobial peptide research, objectives of the thesis and contributions of the thesis. Chapter 2 gives an overview of the field of antimicrobial

peptides and how bioinformatics is facilitating the understanding of AMPs at peptide and gene level (Chapter 1).

Part II describes the implementation of specialized data warehouse of antimicrobial peptides – ANTIMIC integrated with bioinformatics tools (Chapter 3). In-depth usage and sequence analysis done of AMP families using ANTIMIC Profile tool that is integrated in the ANTIMIC database is discussed in Chapter 4.

Part III presents the original findings of the study that includes comparative genomic sequence analysis to find TFBSs by *ab-initio* motif searching approach using Dragon Motif Builder tool in several groups of AMPs (Chapter 5). The findings have led to some important observations about the families of TFs that may potentially regulate AMPcgs.TFBS modules were generated from the promoter analysis of some AMP groups and this provided insights into the concept of conserved TFBS framework in regulation of well-studied and novel AMP groups in Chapter 6. Chapter 7 presents the results of the scan done using the TFBS modules generated in Chapter 6 across human promoter dataset.

Part IV (Chapters 8 and 9) discusses and draws conclusions from the bioinformatics-based approach to large-scale analysis of antimicrobial peptides. It also discusses future directions respectively.

The work presented in this thesis has been published in the following journals,

1) Brahmachary, M., Krishnan, S.P., Koh, J.L., Khan, A.M., Seah, S.H., Tan, T.W., Brusic, V. and Bajic, VB. ANTIMIC: a database of antimicrobial sequences. Nucleic Acids Res. 2004 Jan 1; **32**(Database issue): D586-9.

2) Brahmachary, M., Schönbach, C., Yang, L., Huang, E., Tan, S.L., Chowdhary, R., Krishnan, S.P.T., Lin, C.-Y., Hume, D.A., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. and Bajic, V.B.. Computational promoter analysis of mouse, rat and human antimicrobial peptide-coding genes (*accepted in BMC Bioinformatics*).

Conference presentation

a) A Hybrid Algorithm for Motif Discovery from DNA Sequences (Edward Wijaya, Kanagasabai Rajaraman, Manisha Brahmachary, Vladimir B. Bajic). Poster presented at Asia Pacific Bioinformatics Conference (APBC 2004) held in Singapore.

b) Poster on ANTIMIC database for European Conference of Computational Biology (ECCB 2003, September) held in Paris.

c) Poster on Ab-initio identification of Promoter Elements in Antimicrobial Peptide-coding Genes in 17th International Conference on Genome Informatics, at Yokohama, Japan, December 18-20, 2006.

Part I: Chapter 2: Overview of AMPs

The seat of knowledge is in the head, of wisdom,
in the heart.

(William Hazlitt)

2.1 Properties of antimicrobial peptides

Antimicrobial peptides are ancient weapons of the innate immune system. They are categorized under the first line of defense system of complex higher organisms and probably the only defense system in simpler organisms like bacteria. They are widely present in the animal and plant kingdom. Hence, there are numerous families of these AMPs and new ones are been discovered regularly. They are an effective weapon against an array of pathogens. The antimicrobial peptides intelligently target the microbial cellular membrane and exploit the inherent difference between microbial cell membrane and multicellular plants and animals. They are mostly cationic peptides though there are examples of anionic peptides also which kill pathogens typically by permeabilizing their cell membrane. Interestingly, most pathogens have not been able to develop resistance against them. (Zasloff, 2002).

These cationic AMPs usually have <100 amino acid residues, with at least two positive charges due to lysine and arginine residues and around 50% hydrophobic amino acids (Hancock and Diamond, 2000). There are more than 50 families of AMPs and more than 800 AMPs (Kamysz, 2005). Most AMPs are derived from larger precursors that include a signal sequence. They go through post-translational modifications that include proteolytic processing, and in some cases glycosylation (Bulet *et al.*, 1993), carboxy-terminal amidation and amino-acid isomerization, and halogenation (Zasloff, 2002). Many of these peptides are gene-encoded and synthesized by ribosomes. However, some peptides are derived as cleaved portions from larger proteins, such as buforin II from histone 2A (Park *et al.*, 1996) and lactoferricin from lactoferrin (Bellamy *et al.*, 1992). These peptides are known to be so diverse that the same peptide sequence is rarely

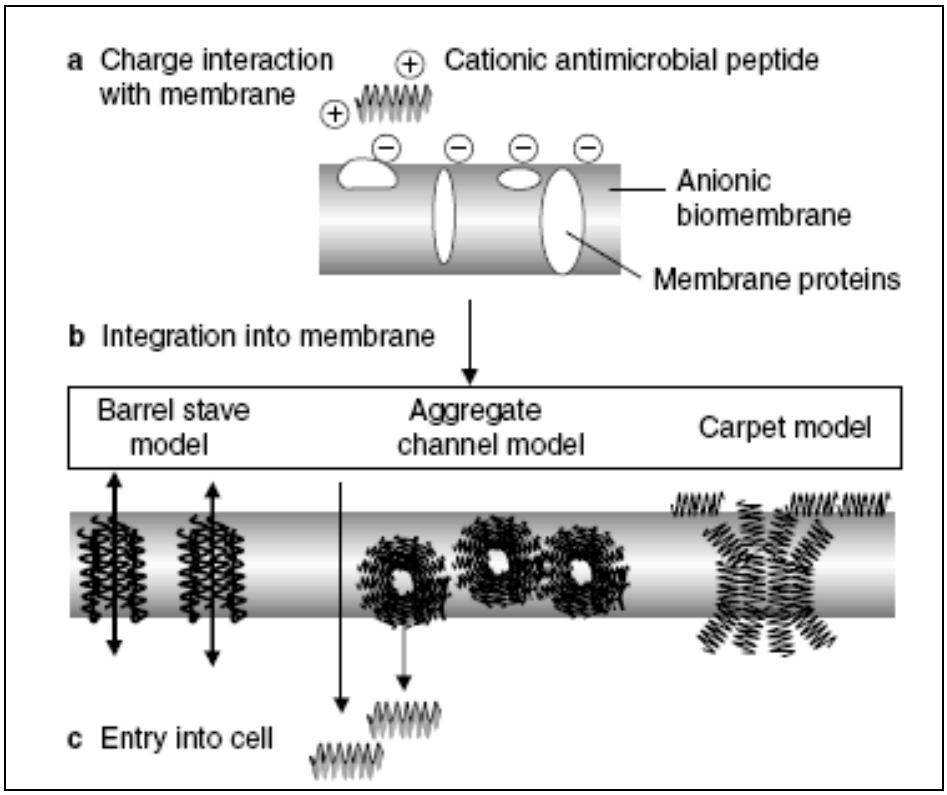
recovered from two different species of animal, even those closely related (Maxwell *et al.*, 2003). Exceptions include peptides cleaved from highly conserved proteins, such as buforin II (Zasloff, 2002). However, within the antimicrobial peptides from a single species, and between certain classes of different peptides from diverse species, significant conservation of amino-acid sequences can be recognized in the pre-proregion of the precursor molecules (Simmaco *et al.*, 1998). This suggests that the pre-proregion is probably conserved, as they are involved in secretion and intracellular trafficking of the peptide. The highly diverse nature of antimicrobial peptides arises from the need of each organism to adapt and survive in different microbial environments. Hence, even single mutations can dramatically alter the biological activity of these peptides (Boman, 2000).

2.2 **Mechanism of action of AMPs**

Antimicrobial peptides act by targeting the membranes of microbes that have a fundamental difference with multicellular animals. In bacterial membrane, the outermost leaflet of the membrane bilayer, which is the exposed surface, is heavily populated by lipids with negatively charged phospholipids head groups. In contrast, the outer leaflet of the membranes of plants and animals is composed principally of lipids with no net charge (Matsuzaki, 1999). Most of the lipids with negatively charged head groups are segregated into inner leaflet, facing the cytoplasm. Shai (1999), Matsuzaki (1999) and Huang (2000) proposed a model for AMP-bacterial membrane interaction (Shai, 1999 , Matsuzaki, 1999, Yang L. *et al.*, 2000). According to the model, the cationic peptides interact electrostatically with the negatively charged membrane. They adopt amphipathic structure where the positively charged residues are lined up on one side and the non-polar residues arranged on the other side of the peptide to be able to accommodate the specific

conditions at the membrane-water interface. This is followed by displacement of lipids, alteration of membrane structure and in certain cases entry of the peptide into the interior of the target cell. Three models have been proposed to describe the molecular events taking place during the peptide-induced leakage of the target cell. **Figure 2.1** is a graphical representation of these models which have been discussed in detail in the following section.

Figure 2.1: Mode of action of AMPs



a) cationic antimicrobial peptide interact with anionic membrane surface and form amphpathic structure. b) pore formation models; the AMPs can integrate into the membrane in three ways barrel stave model, carpet model, aggregate model. Figure has been adopted from (Koczulla and Bals, 2003)

2.2.1 Barrel stave model

According to the barrel stave model after initial electrostatic binding to the outer leaflet of the bacterial membrane, alpha helical amphipathic peptides group together into barrel-like clusters that line amphipathic trans-membrane pores. The non-polar side chains face the hydrophobic fatty acid tails at the inside of the phospholipids bilayer and the hydrophilic side-chains are pointed inward into the water-filled pore. Progressive recruitment of additional peptide monomers leads to a steadily increasing pore size. Leakage of intracellular components through these pores subsequently leads to cell death (van 't Hof *et al.*, 2001).

2.2.2 Carpet model

The carpet model proposes that the AMP clusters cover the surface of the membrane like a carpet. The membrane then collapses at the point of saturation of the concentration of the AMPs. In a short period of time, wormholes are formed all over the membrane leading to an abrupt lysis of the microbial cell. The lipid layer bends back on itself like the inside of a torus. The lateral expansions in the polar head group region of the bilayer are filled up by individual peptide molecules (Shai, 2002). This model has been the proposed mechanism for magainins (Bechinger *et al.*, 1993).

2.2.3 Aggregate Channel model

Another model known as the aggregate channel model proposes that after binding to the phospholipids head groups, the peptides insert into the membrane and then cluster into unstructured aggregates that span the membrane. These aggregates are proposed to have water molecules associated with them providing channels for leakage of ions and possibly larger molecules through the membrane. This model essentially differs from the other two in the way that only short-lived trans-membrane clusters of an undefined nature are formed, which allow the peptides to cross the membrane without causing significant membrane depolarization. Once inside, the peptides proceed to their intracellular targets to exert their killing activities. Another mechanism that has been suggested on AMP-bacterial membrane interactions focuses on self-promoted uptake of AMP (van 't Hof *et al.*, 2001). The cationic peptides bind to the negatively charged LPS present on the surface of Gram-negative bacteria. In the process of binding to LPS, they displace cations like Ca2+ and Mg2+ that are necessary for cell surface stability. This causes disruption in the surface of membrane, and eventually with formation of pores, larger molecules enter the cell. This self promoted uptake pathway works not only in Gram-negative bacteria but also in Gram-positive bacteria (Nykanen *et al.*, 1998).

The ability of AMPs to bind non-specifically to negatively charged membranes and induce pore formation makes them capable of being able to attack a variety of microbes (Gram-positive, Gram-negative bacteria, fungi, virus, and protozoa). However, recently it has been discovered that AMPs also bind specifically to target molecules on the surface of pathogenic membranes to carry out their lytic activities. Nisin binds with high affinity

to Lipid II, the fatty acyl proteoglycan anchor in the bacterial membrane, from which it subsequently diffuses into the surrounding membrane (Brotz *et al.*, 1998). Some plant defensins also use a similar strategy (Thevissen *et al.*, 2000).

After the AMPs bind to the cell surface of the pathogens, many of them do not kill the pathogen merely by permeabilizing the cell membrane. Several of the AMPs have intracellular targets that they bind to and inhibit, thus causing the death of the pathogen. Drosophila AMP, attacin blocks transcription of the *omp* gene in *E.coli* (Carlsson *et al.*, 1991). Bactenecins (Bac5, Bac7) inhibit protein and RNA synthesis of *E.coli* and *Klebsiella pneumoniae* by inhibiting the respiration pathway in addition to permeabilizing their membrane (Skerlavaj *et al.*, 1990). PR-39 has been shown to kill *E.coli* by inhibiting its DNA and protein synthesis (Boman *et al.*, 1993). Neutrophil antimicrobial peptide 2 (eNAP-2) from horse, target and inactivate microbial serine proteases like subtilisin A and proteinase K (Couto *et al.*, 1993).

2.3 Therapeutic applications of AMPs

The short peptide length and versatility of AMPs in targeting a variety of pathogens has generated lot of interest in labs and pharmaceutical industries to create these peptides synthetically and also create hybrids of these peptides to increase efficacy of their functional range (Ferre *et al.*, 2006, Saugar *et al.*, 2006 , Hongbiao *et al.*, 2005). AMPs also seem to be the potential answer to pathogens that have cleverly grown resistant to conventional antibiotics. Most pharmaceutical endeavors have been to develop topical applied agents from AMPs, as the long-term toxicology of these AMPs is not fully understood to facilitate development of safe oral drugs. One such example is magainin

analogue Pexiganan (Ge *et al.*, 1999). Another hurdle is that many of these AMPs show effective pathogen killing *in vitro*, but *in vivo* efficient killing requires high concentration of AMPs that can cause host cell toxicity. **Table 2.1** lists the AMPs that have been commercialized.

Many other applications of AMPs as anti-infective agents have been demonstrated. AMPs have shown potential for being 'chemical condoms' to inhibit the spread of sexually transmitted diseases from pathogens like Neisseria, Chlamydia, human immunodeficiency virus (HIV), Herpes simplex virus (HSV) (Yasin *et al.*, 2000). AMPs in tandem with the conventional antibiotics have shown to increase potency of antibiotics *in vivo* by facilitating access of antibiotics into the bacterial cell (Darveau *et al.*, 1991, Giacometti *et al.*, 2000). LL37 has been tested in animal model to alleviate pulmonary bacterial infection associated with cystic fibrosis (Bals *et al.*, 1999). Medical devices such as intravenous catheters are laced with magainin peptides

that are bound to them by covalent bonds and this facilitates inhibition of microbial colonization and growth on their surfaces (Haynie *et al.*, 1995). AMPs are being used as imaging probes for bacterial and fungal infections due to their specific affinity for microbial membranes (Welling *et al.*, 2000).

Table 2.1: Commercial Development of AMPs

This table has been adopted from (Zasloff, 2002) and modified after (Gordon *et al.*, 2005)

Peptide	Source AMP	Activity	Target disease	Company	Stage
Pexiganan (Msi-78)	Magainin 2	Bacteria	Infected Diabetic Food Ulcers	Magainin (Genaera)	Completed Phase III; not approved by FDA, pending additional studies
Mbi-226	Indolicidin	Bacteria, Fungi	Catheter Sepsis	Micrologix	Phase III
Mbi-594	Cathelicidin-Based, Indolicidin-like	Bacteria	Acne	Micrologix	Phase II, finished
Iseganan (Ib-367)	Protegrin	Bacteria, Fungi	Mucositis	Intrabiotics Pharmaceuticals	Phase II, oral - topical use, failed
P-113	Histatins	Bacteria, Fungi	Oral Candidiasis, Mucositis	Demegen	Phase II
Heliomycin		Bacteria	Antibacterial	Entomed	Preclinical
Human Lactoferricin		Fungi		Am Pharma	Preclinical
Xmp.629	Bpi	Bacteria	Antimicrobial Activity Against *P. Acnes*	Xoma	Phase III
Neuprex (Rbpi21)	Bpi	Bacteria	Reduce Inflammatory Complications Associated With Pediatric Open Heart Surgery Patients	Xoma	Phase I/II

2.4 Regulation of AMP genes

Since AMPs can be both gene encoded peptides and cleaved products, it is likely that their induction and expression fall under numerous different regulatory mechanisms which are yet to be deciphered (Koczulla and Bals, 2003). Some parts of the regulatory mechanisms have been studied in AMPs like beta defensin, alpha defensins in human, mouse and bovine species (Wehkamp *et al.*, 2004, Witthoft *et al.*, 2005, Sherman *et al.*, 2006, O'Neil, 2003, Fang *et al.*, 2003, Musikacharoen *et al.*, 2001, Fehlbaum *et al.*, 2000, Yamamoto *et al.*, 2004). While expression of alpha defensins are generally constitutive (Chen *et al.*, 2006), beta defensin expression in general is induced by different stimuli (Chen *et al.*, 2006) like microbial signals, developmental signals, cytokines, neuroendocrine signals in tissue specific manner. For example hBD-2 expression gets up regulated by infections and inflammatory stimuli (Taguchi and Imai, 2006, Voss *et al.*, 2006, Rivas-Santiago *et al.*, 2005, Kao *et al.*, 2004). Factors like interleukins (IL-1alpha, IL-1beta), tumor necrosis factor-alpha, microorganisms (Gram-positive and Gram-negative bacteria, *Candida albicans*) and LPS are some of the stimulatory agents for expression of beta defensins (Singh *et al.*, 1998, O'Neil *et al.*, 1999, Bals *et al.*, 1999). NF-kB binding site has been found in promoter regions of beta defensins (Diamond *et al.*, 2000). Intracellular signaling probably includes NF-kB, NFIL-6, and JAK/STAT pathways (Kao *et al.*, 2004, Jang *et al.*, 2004). One of the mechanisms of induction of antimicrobial peptides has been deciphered in Drosophila (Imler and Bulet, 2005, Naitza and Ligoxygakis, 2004) and an analogous mechanism exists in humans (Williams, 2001). It has been shown that Toll-like receptors recognize ligands like bacterial LPS and trigger

the signaling cascade that cause induction of some AMP genes (Danilova, 2006). Different signaling cascades are triggered by diverse pathogens in Drosophila. This yields different sets of peptides. For example, the Toll receptor pathway is activated in response to fungi or Gram-positive bacteria while the immune deficiency gene pathway is activated in response to Gram-negative bacteria (Lemaitre *et al.*, 1997, Michel *et al.*, 2001, De Gregorio *et al.*, 2002). However, a lot more needs to be known in terms of the regulatory mechanisms of AMPs.

To understand the regulatory mechanism of AMPs or any other genes, the identification of regulatory elements is the first step. Computational biology can facilitate identification of these regulatory elements faster than experimental identification. Over the years, the growing amount of genomic sequences of different species has facilitated validation and fine-tuning of the computational protocols for transcriptional regulation analysis. The aim is to identify the right transcription factor binding sites in regulatory regions like promoters. Promoters are identified computationally through mapping TSS (Transcription Start Sites) of genes and extracting the upstream regions. Once this data is in hand, it is then possible to search for cis-regulatory elements computationally by screening genomic sequences for the presence of TFBS motifs that have already been identified. TFBSs are usually short (5–25 bp), degenerate sequence motifs that occur very frequently in the genome, hence a position weight matrix (PWM) is often used to quantitatively represent the binding specificity of these factors. More advanced algorithms also facilitate search for pairs or multiple TFBSs in a combination that could be biologically relevant. To reduce the number of false positives, comparative genomics between closely related species is taken into account to find more functionally relevant

TFBSs. Chapters 5, 6 and 7 discuss in details the various current approaches and algorithms that are been used to achieve the above stated objectives.

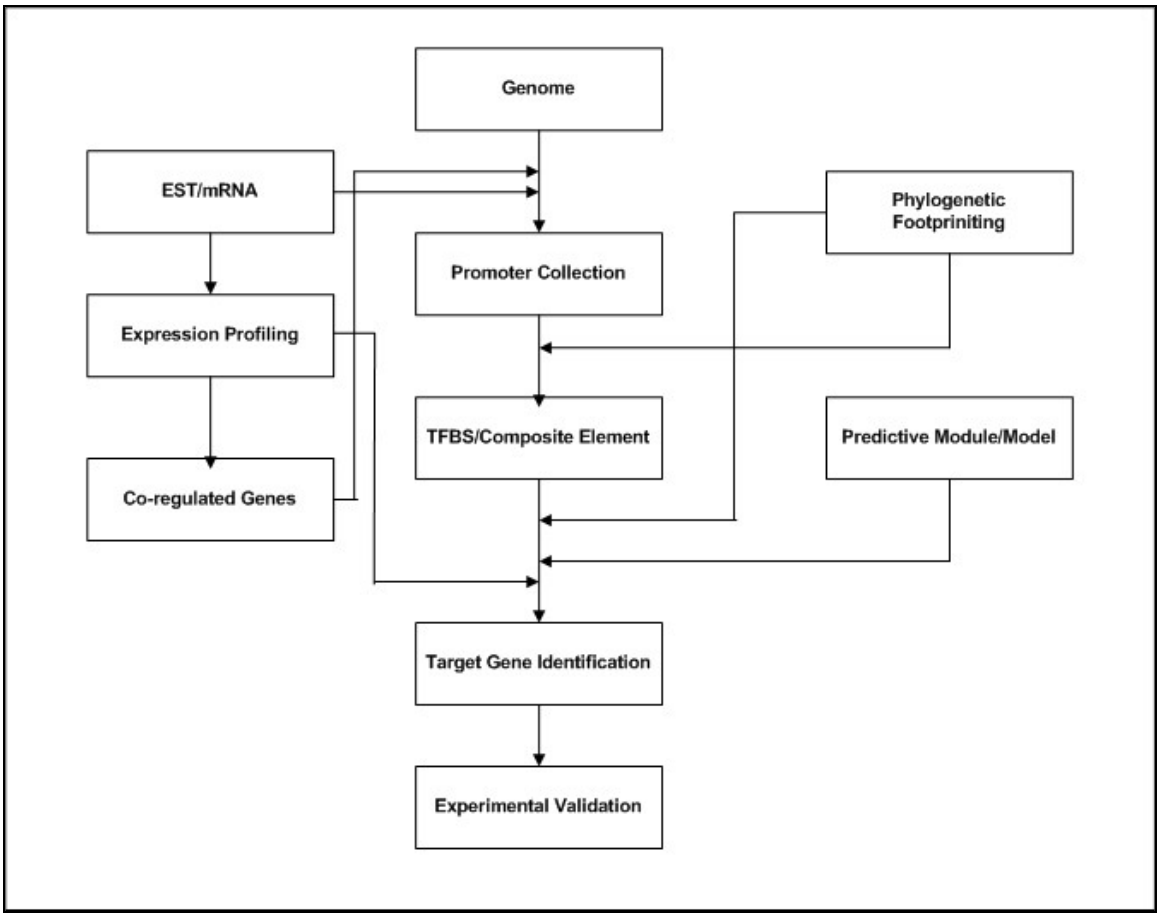
The systematic integration of diverse data types (e.g., individual TFBS hits generated by PWM or IUPAC strings, expression data, sequence data from multiple organisms etc.) together with the development of progressively more sophisticated computational algorithms for promoter prediction, regulatory element identification, and TF coordination modeling, as well as the accumulation of experimental databases of genes and TFs (such as TRANSFAC, TRANSCompel, etc.), will synergistically yield new information and reduce data output to a manageable scale for further experimental validation, thus providing an integrated platform for deciphering the transcriptional regulatory networks.

Figure 2.2 summarizes the general strategy that is implemented computationally in the research of transcription regulatory domain. The starting point is identification of promoter regions using either mRNA/EST mapping or *in silico* promoter prediction (Bajic *et al.*, 2002, Sonnenburg *et al.*, 2006). Co-regulated genes are then derived from expression profiling analysis to refine the promoter dataset to be analyzed. The promoters are subjected to TFBS or composite elements analysis. A predictive regulatory module can be further derived through statistical model building. The module or original TFBS can be used to find other genes regulated in a similar pattern. Comparative genomics (phylogenetic footprinting) can be used both target gene identification and TFBS identification. Expression profiling can also be used to validate the *in silico* target gene prediction. The ultimate test for validity of predictions made by computational methods is still *in vivo* experimental analysis.

In the thesis, a slightly different strategy has been employed, although the essence of the general strategy is retained as shown in **Figure 2.2**. The author has first derived the TFBS modules from computational analysis of AMPcg promoter regions and scanned a larger promoter dataset to find other co-regulated genes. Thus, this study also shows extraction of putative co-regulated genes using computational approach. The co-regulated gene set is then compared to co-expression data derived from expression profiles as a reference to check for the validity of the scanned results.

Figure 2.2: Flowchart of computational analysis for transcriptional regulatory based research

This graphical representation has been redrawn from (Siggia, 2005).



Part II: Chapter 3: ANTIMIC database

One who understands much displays a greater simplicity of character than one who understands little.

(Alexander Chase)

3.1 Introduction

New AMP peptides are being discovered continuously from different organisms experimentally and there is a vast amount of data on natural AMPs but it is not available through one central resource. Bioinformatics facilitates an effective way to store and analyze large volumes of complex biological data through creation of databases. This chapter focuses on resources containing antimicrobial peptide data, the creation of the ANTIMIC database by the author and bioinformatics applications for analysis of antimicrobial peptide data.

3.2 Background

3.2.1 Significance of bioinformatics in antimicrobial peptide research

AMPs are important components of the innate immune system of many species. These peptides are found in eukaryotes, including mammals, amphibians, insects and plants, as well as in prokaryotes (Simmaco *et al.*, 1998, Kylsten *et al.*, 1990, Dangl and Jones, 2001, Luders *et al.*, 2003). Other than having pathogen-lytic properties, these peptides have other activities like antitumor activity, (Kamysz *et al.*, 2003) mitogen activity, or they may act as signaling molecules (Kamysz *et al.*, 2003). Their short length, fast and efficient action against microbes and low toxicity to mammals, have made them potential candidates as peptide drugs (Koczulla and Bals, 2003). In many cases, they are effective against pathogens, which are resistant to conventional antibiotics (Pereira, 2006). They can serve as natural templates for the design of novel antimicrobial drugs (Gordon *et al.*, 2005, Koczulla and Bals, 2003).

Resourceful use of the two approaches (experimental and bioinformatics) can facilitate great strides in understanding the properties and effect of AMPs in biological context. The main goal is the extraction of new knowledge from large-scale analysis of AMP data. The bioinformatics approach provides means for systematic study of a large number of AMPs, and facilitates experimental design and selection of key experiments.

3.2.2 Sources for antimicrobial peptide data and related information

Antimicrobial peptide related data and information can be found across various resources. The data include nucleotide and amino acid sequences, post-translational modifications, secondary structures and 3D structures deposited in public databases such as GenBank (Benson *et al.*, 2005), Swiss-Prot (Bairoch *et al.*, 2004) and PDB (Deshpande *et al.*, 2005). Structure-function information, and mutation studies data, is available in the literature. The advantages and disadvantages of these databases for the creation of a database of antimicrobial peptide will be reviewed in the next sections. The issues of data collection, cleaning and annotation when consolidating the scattered data have also been described.

3.2.2.1 GenBank and GenPept databases

Antimicrobial peptide data are extracted from GenPept protein database of GenBank (Benson *et al.*, 2004) which contains publicly available translated nucleotide sequences found in GenBank. GenBank stores data that are direct submissions from labs and batch submissions from large-scale sequencing projects to help maintain accuracy, relevance and comprehensiveness of the database. However, records in these databases contain only

basic information such as the AMP sequence, its name, taxonomy of the source organism, and when available, a list of basic sequence features and references. Also, there are many instances of entries of partial peptide sequences being present though a different entry contains the whole peptide sequence of the same gene. Hence, a certain amount of redundancy remains in the Genbank database. The records need to be enriched with structural and functional information that is available in literature.

3.2.2.2 Swiss-Prot and TrEMBL databases

Swiss-Prot and TrEMBL (Bairoch *et al.*, 2004) databases were also used as resources for extraction of data for creation of ANTIMIC database. These databases have a comprehensive collection of annotated protein sequences. They contain structural and functional information about peptide sequences that may include disulfide connectivity, information on secondary structure and protein family classification, among others. The information in the records expedites subsequent annotation when new structure-function information is available.

3.2.2.3 Protein Data Bank (PDB)

Analyzing antimicrobial 3D structures are important because function is related to its structural folding. Inclusion of 3D structural information to antimicrobial sequence analysis facilitates identification of residues that are important for structure and function. This in turn aids the process of designing effective synthetic antimicrobial peptides.

3.2.5 Issues on data collection, cleaning, annotation

Different databases have different formats and variations in fieldnames that describe the same information. This poses problems when data needs to be extracted in an automated manner from these sources.

For example, an AMP primary sequence is described in the 'translation' field of a GenBank record but in Swiss-Prot, it is described in the 'sequence' field. Standardization of data representation across different databases will definitely enable a smoother extraction process and cross-referencing of fields across different databases. For example, a standard field such as 'translation' can be used to describe AMP primary sequence regardless of data sources. The uniform data representation is critical because consistency is required for efficiency of subsequent analyses.

When consolidating records from different databases, the same data may be duplicated in another database, resulting in data redundancy. Data cleaning involves removing these redundant records to improve on data quality. Data cleaning also involves detecting discrepancies in data information, highlighting, and subsequently correcting the conflicts.

Records in the public databases typically contain basic information. Data annotation, also known as data enrichment or enhancement, is the process of furnishing critical commentary or explanatory notes. Data annotation enriches the data for extrapolation of meaningful insights from multi-source bits of information. Correlating the relevant information from multiple sources is critical for increasing the overall knowledge and understanding of a specific subject in the data warehouse (Karasavvas *et*

al., 2004). It is important to differentiate experimentally determined function from those that have been predicted computationally (Karp *et al.*, 2001) because the latter require subsequent validation. This would allow researchers to verify and decrease the propagation of incorrect predicted function during data annotation.

3.2.4 Data warehouses of antimicrobial peptides

To the author's knowledge, four antimicrobial databases exclusive of the author's database (ANTIMIC) are currently available as major resources for the study of antimicrobial peptides. These meta-databases (databases for storing metadata (data that describes data) for a specific purpose) contain entries collected from different sources.

An attempt has been made in Italy to consolidate information about AMPs and store it in a database called AMSdb (http://www.bbcm.univ.trieste.it/~tossi/search.htm). This database contains annotated AMP sequence data and enables a keyword search for categories such as ID, date, family, category, activity, organism source, and generic keywords. The AMSdb database consists of 804 entries (as of 05 August 2003) of eukaryotic origin only. This database does not provide any tools for the analysis of data.

Another database (http://public-1.cryst.bbk.ac.uk/peptaibol/home.shtml), Peptaibol database, is a highly specialized one that contains over 300 entries of antibiotic peptides known as Peptaibols (Chugh and Wallace, 2001), that originate from fungal organisms like *Trichoderma* and *Emericellopsis*. This database enables users to search for information about Peptaibols by name, or Peptaibol group. It also allows for searching of entries using motifs specific for Peptaibols (that are known to have non-standard amino acid residues in them). The database stores Peptaibol entries with PDB entries and

enables users to view the structure from the database. The authors of this database have classified the Peptaibols into subfamilies based on the alignments of these sequences with common sequence features thought to be important for channel formation (Chugh and Wallace, 2001).

The APD (the Antimicrobial Peptide Database) is another data resource for AMPs. It contains annotated information for 559 peptides (498 antibacterial, 155 antifungal, 28 antiviral and 18 antitumor, some peptides are member of multiple groups). It has an interactive interface for peptide query, prediction and design. It also provides statistical data for a select group of or all the peptides in the database. Peptide information can be searched using keywords such as peptide name, ID, length, net charge, hydrophobic percentage, key residue, unique sequence motif, structure and activity. APD facilitates studying the structure–function relationship of antimicrobial peptides. The database can be accessed via a web-based browser at the URL: http://aps.unmc.edu/AP/main.html (Wang and Wang, 2004).

SAPD (Synthetic Antibiotic Peptides Database) (http://oma.terkko.helsinki.fi.8080/~SAPD) contains information about peptide antibiotics that have been that have been synthesized based on naturally occurring structures of antimicrobial peptides. This database caters to researchers who want information about the various structure manipulation experiments that have been done on AMPs. It contains only 22 entries of synthetic peptides with detailed information.

Compared to these databases, ANTMIC contains the most number of AMPs that cover sequences from both eukaryotes and prokaryotes. It has sequence analysis tools integrated with the database that are unique to this database. These tools are

sequence similarity search tool such as BLAST, a peptide structure viewer tool to view the 3-D peptide structure and analytical tools like the Antimic profile module, facilitate analysis and classification of AMPs. The details of these tools have been discussed in the following sections.

Table 2.2: Comparison of the various antimicrobial peptide databases

Database	Data	Tools	Address	No. of sequences
ANTIMIC	eukaryotic and prokaryotic peptide sequences AMP peptide structures	BLAST(peptide), HMM based peptide sequence profiling tool for sequence – function relationships of AMPs	research.i2r.a-star.edu.sg/Templar/DB/ANTIMIC/	1788
AMSdb	eukaryotic peptide sequences AMP peptide structures		http://www.bbcm.univ.trieste.it/~tossi/search.htm	804
Peptaibol database	peptaibol AMPs only peptaibol peptide structure	sequence based classification	http://public-1.cryst.bbk.ac.uk/peptaibol/home.shtml	317
APD	eukaryotic peptide sequences	tools for studying structure–function relationship of AMPs	http://aps.unmc.edu/AP/main.html	559
SAPD	peptide antibiotics synthesized based on naturally occurring structures of AMPs		http://oma.terkko.helsinki.fi.8080/~SAPD	22

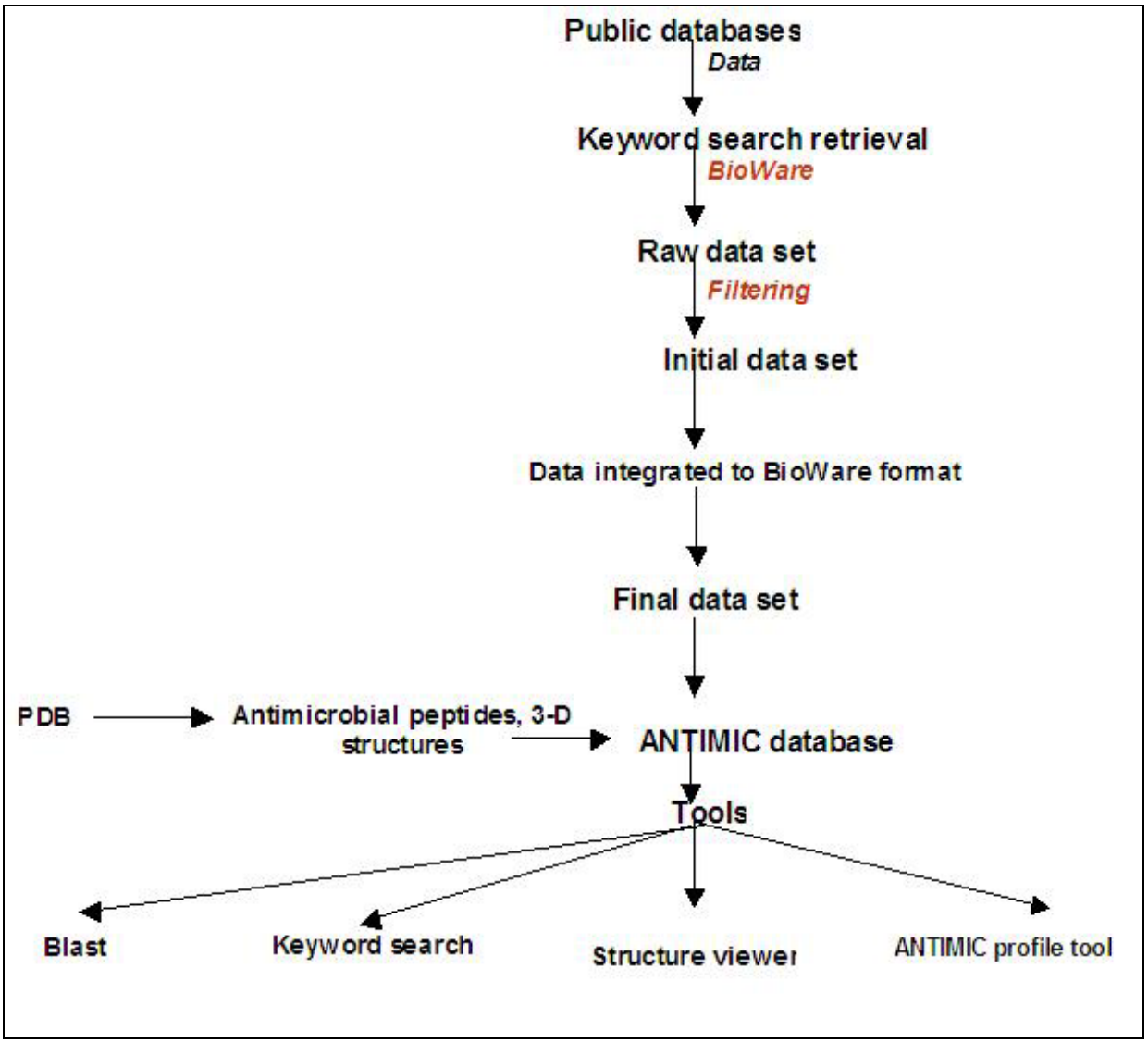
3.2.5 Bioinformatics tools

The next step after creating a comprehensive collection of data and storing it in databases is the use of computational tools for analysis to extract biologically meaningful information. General bioinformatics tools commonly used in sequence analyses of antimicrobial peptide data include but are not limited to BLAST (Altschul *et al.*, 1997) and Clustal W (Thompson *et al.*, 1997). The BLAST search tool finds regions of local

similarity between query sequences and database sequences by calculating the statistical significance of matches. Uses of BLAST include inferring functional and evolutionary relationships between sequences as well as helping to identify members of gene families. Clustal W is a general purpose multiple sequence alignment program for nucleotide or protein sequences. It involves the optimal alignment of the greatest number of identical or similar residues into columns across many nucleotide or protein sequences. Patterns of aligned sequences can be used in the analysis of function, structure and phylogeny relationship between sequences.

3.3 Materials and methods

Figure 3.1: Methodology for building the ANTIMIC database



The ANTIMIC database contains an extensive collection of antimicrobial sequences from many families. The database has been created on an in-house data-warehousing platform

(BioWare, sdmc.i2r.a-star.edu.sg/Templar) that enables building of specialized searchable biological databases. BioWare comprises three program modules: BioWare Retreive Module retrieves raw data from diverse sources on the internet; BioWare-Prep Module processes retrieved data, and Templar Module integrates this information into a central repository. The processing includes generation of a report summary for removal of redundant entries, renumbering of entries, and other sub-modules like a module for generation of multiple alignments and a module for viewing cysteine bridge patterns to help the database creator to manage the information more efficiently.

4.3.4 Data collection for the ANTIMIC database

The data has been extracted from public databases. Specific keyword search terms like “alpha defensin” and generic keyword terms like “antibacterial”, “antifungal”, etc. were used within the BioWare Retrieve Module to search the NCBI’s GenBank and Swiss-Prot databases.

3.3.2 Data filtering

This preliminary data set was checked for duplicates and redundancies, with help of BioWare and manual curation. Entries that may have been the earlier versions of another entry were removed. This was facilitated by the BioWare-Prep Module that generates a sequence comparison report summary based on pair wise alignment of entries in the data set. Entries, that had 100% sequence identity, were reported as duplicates. Duplicates having the same name and taxon (organism source) were compared. The entry judged to contain the most complete information was kept while the others were not considered.

Duplicates originating from different source species were kept as separate entries. Sequences that shared fragment or partial identity, where one sequence was an identical fragment of another, were checked for their uniqueness by referring to both literature and the cross-references field from the public databases. Most of these entries were earlier versions of other entries in the data set and hence were deleted. All deleted entries were added to the ANTIMIC file of rejected entries (FRE), which is used to avoid future retrieval of the same entry during database updates. This resultant data set will be referred to in this text as the preliminary cleaned data set.

Next, each of the entries were checked manually to ensure that they are the AMP entries and not irrelevant entries, examples including "Integrin" or "Reticulon 4 receptor precursor" which may have been picked up by the keyword search. Records eliminated at this step were recorded in the FRE. The final cleaned data set was used as the input to the Templar Module, and the online version of the ANTIMIC database was generated (http://research.i2r.a-star.edu.sg/Templar/DB/ANTIMIC/).

The antimicrobial sequences were formatted into a blastable database and integrated to the ANTIMIC database.

3.3.3 Antimicrobial structural data incorporation

The ANTIMIC database has a structure viewer module that contains the PDB structures of antimicrobial sequences. The structure viewer was populated by searching the PDB database for 3-D structures of antimicrobial sequences present in the ANTIMIC database. PDB accession numbers present in the annotation of entries in the ANTIMIC database were linked to their corresponding 3-D structures.

3.3.4 Creating the ANTIMIC profile tool for antimicrobial classification

A web-based antimicrobial peptide analysis tool named as ANTIMIC profile tool was created with the aim to facilitate tentative classification of query sequences into different antimicrobial families. It has three modules based on the HMMER software package (http://hmmer.janelia.org/), (Eddy, 1995). The ANTIMIC profile tool uses predefined antimicrobial-specific library of profiles, although users can generate profiles out of their specific sequences. The profile library has been created out of mature peptide regions of AMPs of different families. The mature peptide region from each of the AMP sequences of a family were extracted based on the annotation provided for the AMP sequence in Swiss-Prot or GenBank. The mature domains were then subjected to multiple sequence alignment using Clustal W (Thompson *et al.*, 1994). The Clustal W output was further processed with hmmbuild (part of HMMER package) to create an HMM profile. This profile was stored as one of profiles of antimicrobial-specific library e.g. mellitin.hmm, protegrin.hmm etc.

The 'Query profile' module of ANTIMIC tool was built by first constructing an antimicrobial HMM database. The HMM database was built by concatenation of single HMM profile files created from different AMP family sequences. The files were concatenated using the –A "append" option of hmmbuild program. Then the hmmcalibrate program was run to determine appropriate statistical significance parameters for a HMM prior to doing database searches.

After the HMM profile database was built, known antimicrobial query sequences were used to test the specificity of the HMM profiles in the database. Query sequences that do not have known antimicrobial function were also used as a negative test set (see chapter 4) to check the specificity of 'Query profile' program.

The Query db module is based on the hmmsearch program of HMMER package. For the "Query db" module, the nr dataset from Genbank was downloaded in fasta format and the ANTIMIC database peptide sequences were downloaded in fasta format and these datasets were built into two different databases to be queried against by the HMM profiles created. Details of the various HMMER command options can be found in the HMMER userguide. **Figure 3.1** summarizes the strategy to build the ANTIMIC database.

3.4 ANTIMIC database features

The ANTIMIC database is the most comprehensive source of natural AMPs to date that has been manually curated. The database currently has 1788 number of entries (last updated on October 2002) extracted from GenBank and Swiss-Prot. The entries come from both eukaryotic and prokaryotic organisms. The creation of the database is a systematic collection of AMP sequences and the first step in the computational approach to understand the transcriptional regulation of AMPs. Hence, the intention to create the database was to aid molecular analysis of AMPs. In addition to comprehensive peptide information and AMP specifics, ANTIMIC database has integrated data extraction tools, sequence similarity search tools, BLAST, peptide structure viewer tool, and analytical

tools like the ANTIMIC profile module, all of which facilitate analysis and classification of AMPs. **Figure 3.2** gives the statistics of the data stored in ANTIMIC database.

Figure 3.3 shows the distribution of ANTIMIC data based on structures of various AMP groups. As discussed earlier AMPs have highly variable primary sequences, and they show some degree of conservation at structure level. Thus AMPs have been also classified based on their common structures (van 't Hof *et al.*, 2001, Vizioli and Salzet, 2002). A striking conservation is observed of AMPs that have disulfide bridge**s.** For eg. defensins have been grouped as alpha, beta, theta based of their disulfide architectures (Chen *et al.*, 2006).

3.4.1 Database Organization

Each ANTIMIC entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, bibliographic references, and a table of features listing areas of biological significance, coding regions, peptide regions, sites of mutations or modifications and the protein translation.

The annotation of each entry in the database contains the following fields (**Figure 3.4**): A unique accession number 'DBACC' that defines each record in the ANTIMIC database. The format is (D) (six digit number), where D denotes an entry of AMP and the six-digit number is a unique descriptor of the entry. Next, the field 'Date' identifies the date when the entry was made. The field 'Locus Name', 'Sequence length', and 'GenBank Division' contain information on the locus, length of the sequence, and the division group to which the sequence belongs in GenBank. In some entries 'GenBank Division' is also known as 'Molecular type'. The date when the entry was updated by

public database is shown in the field 'GenBank Modification Date' or, in some entries, as 'Release Date'. The 'Name' field contains the name of AMP used in literature, and if available, their common names. The field 'Accession' provides hyperlinks to the corresponding entries of the relevant external databases, GenBank and Swiss-Prot. The organism source of AMPs can be found in the 'Source' field and its taxonomy is shown in the 'Species' field. The 'Reference' field contains the literature references, with the author names and titles. Relevant comments or observations can be found in the 'Comment' field. Structural features of AMPs, such as residues forming the disulfide bridges, helices or strands, are described in the field 'Features'. Putative structural information derived by similarity to known structures is indicated as 'By Similarty'. Many entries have the field 'Link' that links that entry other databases EMBL, Pfam, and ProDom etc. The field 'Translation' provides the amino acid sequences of an AMP entry. If the PDB structure is available, the field 'Structure' contains internal hyperlinks to the PDB structure stored in ANTIMIC database for relevant records.

3.4.2 Integrated Tools

The ANTIMIC database contains several integrated tools to help in the data extraction and analysis of AMP sequences. The data extraction and sequence viewing tools include:

- Keyword search,
- BLAST search, and
- Structure viewer.

The Keyword search feature allows users to search the database using keywords. The BLAST (Altschul *et al.*, 1990) search enables users to perform sequence similarity

search against the antimicrobial sequences stored in the database. The structure viewer allows for the 3-D structures of individual AMPs to be viewed.

The analysis-based tools consist of the ANTIMIC profile tool. This tool has multiple modules. The modules allow for building of new profiles, querying new sequences against the build profiles or against the predefined profile library, as well as against either ANTIMIC or nr databases. A detailed analysis done using ANTIMIC profile tool is discussed in Chapter 4.

Users can access ANTIMIC entries by using either a simple keyword search such as species name, type of antimicrobial activity, Swiss-Prot or GenBank accession numbers, etc., or they can perform complex searches for more specific results by using more than one keyword with the support of Boolean operators. For example, a simple search would be to use a keyword like “Protegrin” to retrieve entries of this family. A complex search would be “mellitin and wasp” which will return mellitin family related entries that are specific to the wasp species. Therefore, any term that is present in the annotation of the entries can be used in combination with others to retrieve specific results. The results are displayed in a tabular form as a list. The list displays accession numbers, species from which AMP originates, and the antimicrobial sequence name. The accession number is hyperlinked within the database to the full data record.

The database has integrated the BLAST program (Altschul *et al.*, 1990) that consists of a set of similarity search programs for protein or DNA sequences. The BLAST feature allows users to perform sequence comparison using the BLAST algorithm. A query sequence of amino acid can be compared against all sequences in the ANTIMIC database. Users can choose to return the results either in standard BLAST

output or color-coded multiple sequence alignment generated by MView program (Brown *et al.*, 1998). MView highlights the positions of conserved and homologous amino acids in the multiple sequence alignment returned by BLAST.

For the antimicrobial sequences that have an entry in the PDB the corresponding peptide structures can be seen through the structure feature using the Chime (Horton, 1999) or Swiss PDB viewer (Guex and Peitsch, 1997). The PDB files can also be downloaded. The latest version of Chime 2.6 SP4 is functional with Netscape Navigator (version 4.x) (**Figure 3.5**).

3.5 Future work

A database for antimicrobial peptides can prove to be very useful for scientists in academics and commercial organizations. For the ANTIMIC database to be consistently useful to the researchers, data enrichment with more data information on AMPs and regular updating is important. The ANTIMIC database can be further enriched by adding gene information, promoter sequences, gene information, transcript information, gene structure, orthologs and paralogs and gene ontology of known AMPs that have been extracted in the process of understanding AMP regulation. The peptide information can also be enriched in many ways. Some suggestions are peptide cleavage information, amino acid post translation modification of the peptides, known mechanism of action for AMPs and minimal inhibitory concentration (MIC) that can be appended to the existing AMP entries.

New AMPs are being discovered regularly and hence the ANTIMIC database needs to be updated regularly. A semi-automated process with the help of BioWare is proposed to make regular updates possible. Keywords (AMP gene names, family names) of new

AMPs can be collected. BioWare can be used to search for annotation of these AMPs using the keywords and retain only the new entries. Secondly, the ANTIMIC profile tool can be facilitated to extract new AMP sequences based on profile searches.

3.6 Conclusion

ANTIMIC is a specialized database that has been built with the aim of making a comprehensive repository of natural AMPs complemented by data extraction and analysis tools to help further analysis of AMPs. One of the integrated tools, the ANTIMIC profile module, enables users to assign a new putative antimicrobial sequence to a family and functional domain. It also enables the capture of new peptide homologs from other public databases. Chapter 4 gives a detailed view of the utility of the ANTIMIC tool to datamine useful information on antimicrobial peptides.

Figure 3.2: Number of AMP entries in ANTIMIC database in terms of different species

The ANTIMIC database has 1788 entries as of (June 2003)

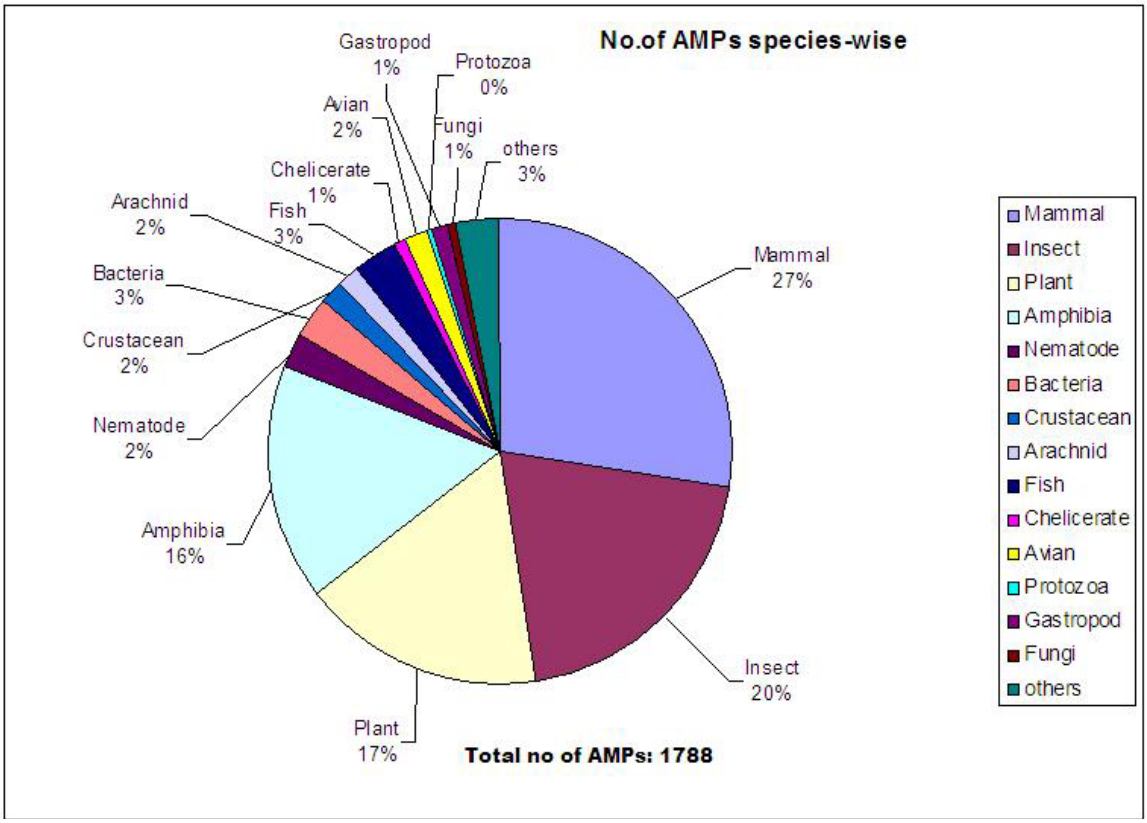


Figure 3.3: Number of AMP entries in ANTIMIC database in terms of different sequence properties

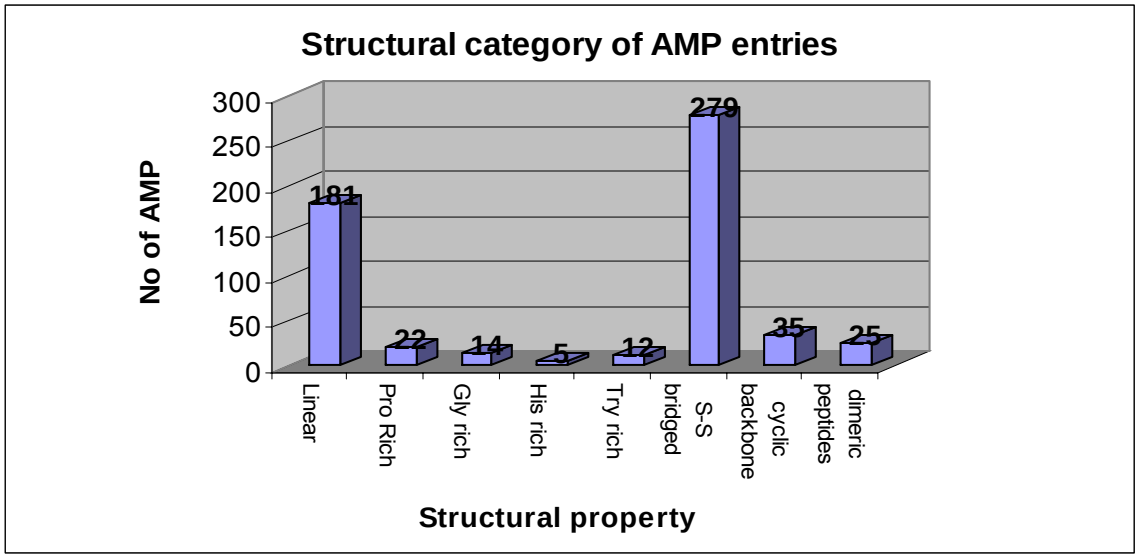
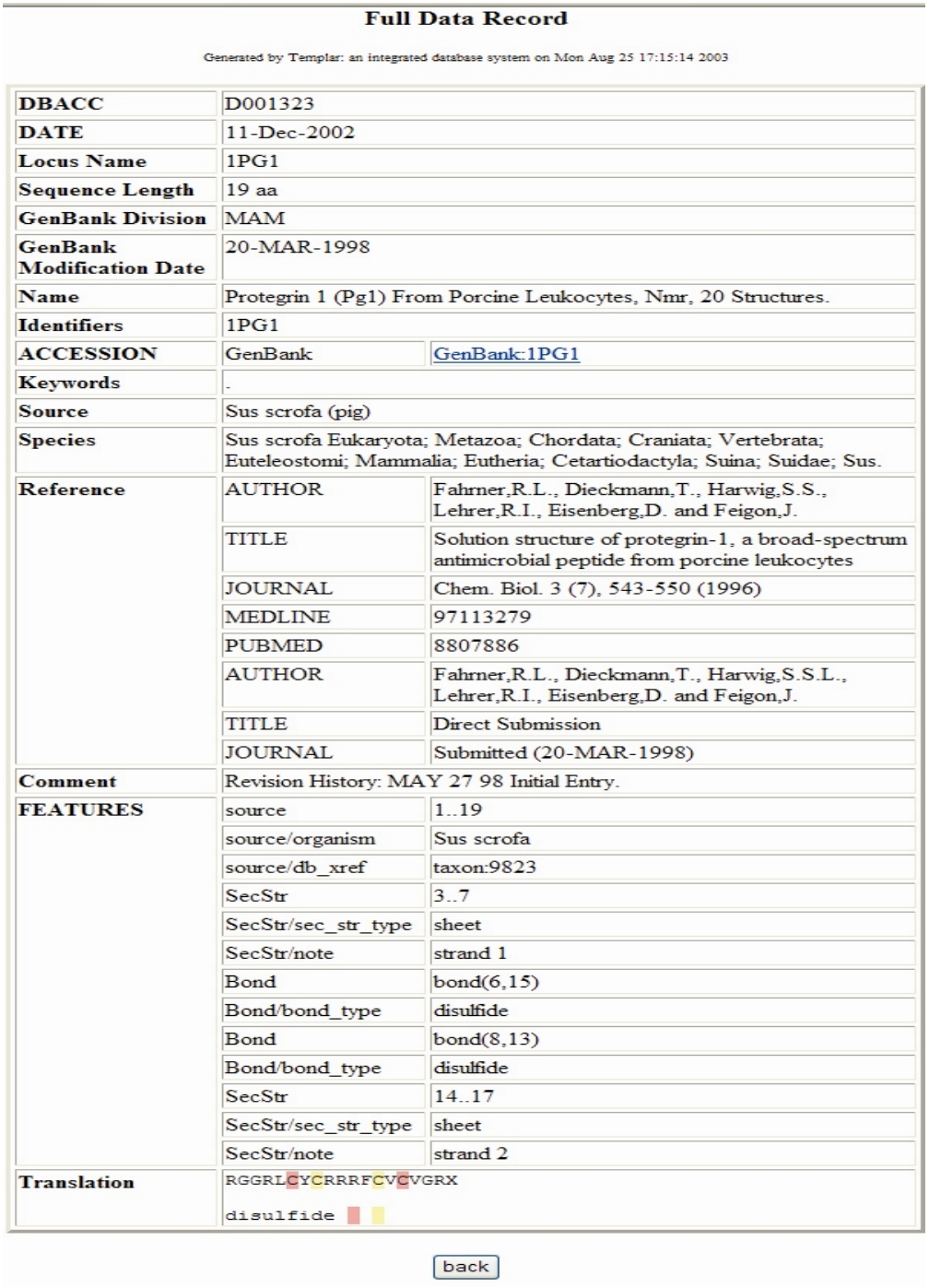


Figure 3.4: A typical ANTIMIC entry



Full Data Record

Generated by Templar: an integrated database system on Mon Aug 25 17:15:14 2003

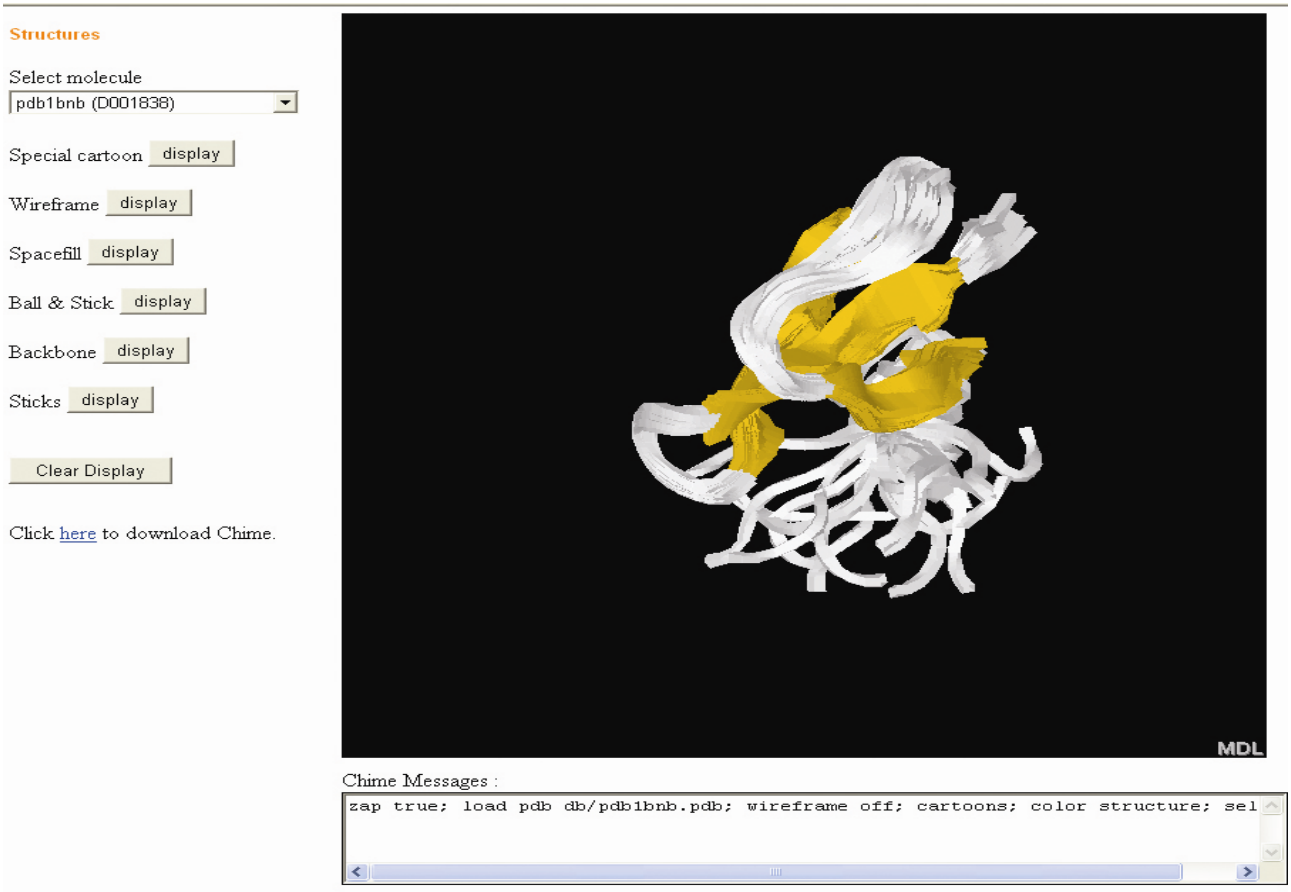
DBACC	D001323	
DATE	11-Dec-2002	
Locus Name	1PG1	
Sequence Length	19 aa	
GenBank Division	MAM	
GenBank Modification Date	20-MAR-1998	
Name	Protegrin 1 (Pg1) From Porcine Leukocytes, Nmr, 20 Structures.	
Identifiers	1PG1	
ACCESSION	GenBank	GenBank:1PG1
Keywords	.	
Source	Sus scrofa (pig)	
Species	Sus scrofa Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Cetartiodactyla; Suina; Suidae; Sus.	
Reference	AUTHOR	Fahrner,R.L., Dieckmann,T., Harwig,S.S., Lehrer,R.I., Eisenberg,D. and Feigon,J.
	TITLE	Solution structure of protegrin-1, a broad-spectrum antimicrobial peptide from porcine leukocytes
	JOURNAL	Chem. Biol. 3 (7), 543-550 (1996)
	MEDLINE	97113279
	PUBMED	8807886
	AUTHOR	Fahrner,R.L., Dieckmann,T., Harwig,S.S.L., Lehrer,R.I., Eisenberg,D. and Feigon,J.
	TITLE	Direct Submission
	JOURNAL	Submitted (20-MAR-1998)
Comment	Revision History: MAY 27 98 Initial Entry.	
FEATURES	source	1..19
	source/organism	Sus scrofa
	source/db_xref	taxon:9823
	SecStr	3..7
	SecStr/sec_str_type	sheet
	SecStr/note	strand 1
	Bond	bond(6,15)
	Bond/bond_type	disulfide
	Bond	bond(8,13)
	Bond/bond_type	disulfide
	SecStr	14..17
	SecStr/sec_str_type	sheet
	SecStr/note	strand 2
Translation	RGGRLCYCRRRFCVCVGRX disulfide	

back

Full data record of AMP named 1PG1 from the protegrin family. Information about cysteine bridges is shown in color for simpler viewing.

Figure 3.5 Structure viewer image

Structure viewer image shows the PDB structures of AMPs. The structures can be viewed with the Chime program that is compatible with Netscape Navigator. The figure shows the structure of beta defensin, (BNDB-12) from *Bos taurus*.



Part II: Chapter 4: HMM based sequence analysis of AMPs

Every artist was first an amateur
(Ralph Waldo Emerson)

4.1 Introduction

In this chapter the author demonstrates the usage of the ANTIMIC profile tool that is integrated with the ANTIMIC database (Chapter 3). ANTIMIC profile tool can help to identify the plausible important residues for AMP peptides that are involved in their antimicrobial function. This tool can also facilitate in assigning new putative AMP sequences to AMP families based on HMM profile matches. It also facilitates search for new AMP sequences based on the profiles of different AMP families.

4.2 Background

4.3.4 Classification of AMPs based on sequence properties

Attempts have been made to classify the huge and diverse collection of AMPs based on biochemical and structural features. The largest number of AMPs are cationic molecules. Based on structural features, cationic peptides are divided into three classes (**Table 4.1**). The first class consists of linear peptides forming alpha-helical structures. Examples under this class comes from cecropins originating from insects which are a family of 3-4kDa linear amphipathic peptides, magainins whose source is from frogs, and cleaved product of histone molecules (buforin II). The second class consists of cysteine-rich open ended peptides containing one or more disulfide bridges. Defensins, which originate from different mammalian species fall under this category. Defensins themselves are arranged in families based on structural differences. The third class comprises peptides rich in specific amino acids such as proline, glycine or histidine. AMPs like drosocin, metchnikowins from Drosophila are proline rich peptides. Attacins and diptericins are glycine rich peptides (Vizioli and Salzet, 2002).

There is a novel group of AMPs isolated from mammalian epithelia, which are anionic in nature (**Table 4.2**). The first class comprises of phosphorylated compounds like peptide B, enkelytin that are cleavage products of neuropeptide precursors like proenkephalin-A. The other group of anionic peptides are aspartic acid rich peptides like dermicidin. At present their mode of action is not clearly deciphered.

There are also aromatic dipeptides that are AMPs. These are low molecular weight antibacterial compounds. Examples of these are N-beta-analyl-5-S-glutathionyl-3,4-dihydroxyphenylalanine identified in *Sarcophaga peregrina.*Another class of anionic peptides are derived from oxygen binding proteins like lactoferrin from human and hemocyanin derived peptide from shrimp. Bactericidal activity of anionic peptides, aromatic peptides and oxygen derived peptides are weak compared to cationic peptides.

Table 4.1: Classification of cationic AMPs.

Reference (PMID)*: Pubmed Unique ID

Structure and representative peptides	Organism	Antimicrobial activity	Mode of antimicrobial activity	PDB structure ID	Reference (PMID)*
Linear alpha-helix peptides					
Cecropins	Insects, pig	Bacteria, fungi, virus,	Carpet		11807545, 10426426, 10333735
Clavanin, styelin	Tunicates	Bacteria			10333735
Magainin, dermaseptin	Amphibians	Bacteria, protozoa	Torroidal pore	2MAG (magainin)	11807545, 10333735
Buforins	Amphibians	Bacteria, fungi	Binds nucleic acids		11807545, 8573171 , 9514864
Andropin	Insects	Bacteria			1899226
Myeloid antibacterial peptide 27 (mature peptide)	Bovine	Bacteria, fungi			8910461
Antibacterial peptide BMAP-34 (mature peptide)	Bovine	?			9409740
Linear peptides rich in certain amino acids					
Pro-rich:					
Drosocin, metchnikowin,	Fruit fly	Bacteria	Inhibits enzymatic activity	1MYN (Drosocin)	10426426
Pyrrhocoricin,	Hemipteran	Bacteria, fungi			10426426
Metchnikowin					
Bactenecin 5 (mature peptide region)	Sheep	Gram-negative bacteria			10417180
Abaecin	honey bee	Gram-negative,Gram-positive bacteria			7961803
Gly-rich:					
Diptericins, attacins	Dipterans	Bacteria			10426426
His-rich:					
Histatin	Human	Bacteria, fungi	Inhibits enzymatic activity		11807545, 10333735
Piscidins	Fish	Gram-positive, Gram-			11739390

		negative bacteria			
Tyr-rich:					
Indolicidin	Cattle	Bacteria	Alters cytoplasmic membrane septum formation, inhibits protein synthesis	1G8C	11807545, 10333735
Single disulfide bridge					
Thanatin	Hemipteran	Bacteria, fungi		8TFV	11807545, 10426426, 10333738
Cyclic dodecapeptide precursor (Cathelicidins)	Bovine	Bacteria			8706679
Brevinins	Frog	Bacteria			11807545, 10333735
Brain natriouretic peptides	Human	Bacteria, Fungi			11410403
Two disulfide bridges					
Tachyplesin II	Horseshoe crab	Bacteria, fungi, virus	Binds nucleic acids		11807545, 10333735, 10333738
Androctonin	Scorpion	Bacteria, fungi		1CZ6	11807545, 10333738
Lactoferricin	Human	Bacteria		1lfc	1599934
Protegrins	Pig	Bacteria, fungi	Torroidal pore	1lxe,1kwy	
Three disulfide bridges					
Alpha defensins	Mammals	Bacteria, fungi	Inhibit protein synthesis (human alpha defensins)		11807545, 10333735
Beta defensins	Mammals	Bacteria, fungi		1IJV;1E4S;1KJ5;1FD3;1FD4;1E4Q;1FQQ;1KJ6;1KJ6;1E4T;1E4R; 1bnb	11807545, 10333735
Defensins	Insects	Bacteria, fungi, protozoa			10426426, 10333738
Penaeidins	Shrimp	Bacteria, fungi			10333738, 11598107
Saposin c-like pore forming peptides	Entamooeba dispar	Bacteria			10518795
More than three disulfide bridges					
Tachycitin	Horseshoe crab	Bacteria, fungi		1DQC;	10333738

Drosomycin	Fruit fly	Fungi		1MYN;	10333738
Gambicin	Mosquito	Bacteria, fungi, protozoa			11606751
Heliomicin	Lepidopteran	Bacteria, fungi		1I2U;	11580275
Defensins	Plants	Fungi			11807545, 10333739

Table 4.2: Classification of non-cationic AMPs.

This table has been modified from Vizioli and Salzet, 2002)

Structure and representative peptides	**Organism**	**Antimicrobial activity**	**Reference (PMID)***
I. **Anionic peptides**			
a. **Neuropeptide derived:**			
Enkelytin	Bovine, human,	Bacteria	11192590, 11377277
Peptide B	Bovine, human,	Bacteria	11192590, 11377277
	Leech, mussel		
b. **Aspartic acid rich:**			
H-GDDDDDD-OH	Ovine	Bacteria	8552650
Dermcidin	Human	Bacteria	11694882
Glu-rich			
Maximins 3/H5	Toad	Gram-positive bacteria	11835991
II. Aromatic dipeptides			
N-β-alanyl-5-S-glutathionyl-3,4-dihydroxyphenylalanine	Flesh fly	Bacteria, fungi	8662858
p-Hydroxycinnamaldehyde	Saw fly	Bacteria, fungi	9923603
Peptides derived from oxygen-binding			
III. Proteins			
Hemocyanin derived	Shrimp	Bacteria	11598107
Hemoglobin derived	Tick	Bacteria	10464258
Lactoferrin	Human	Bacteria, virus	11431038

4.2.2 Computational classification methods

In recent years, many computational approaches and tools have evolved to classify peptides in various ways. Hidden Markov Model (HMM) is one such approach besides Support vector Machine (SVM), Artificial Neural Networks (ANN), Decision trees and so on (Jia *et al.*, 2006). HMM performs better than other machine learning methods in classifying proteins in family and superfamilies (Can *et al.*, 2004). HMM is used to provide statistical representation of real biological processes. One example is classification and characterization of protein families (Bateman *et al.*, 1999). HMM generates optimum multiple sequence alignment for a given protein family that can be used as a method for classification of protein sequences. HMMs can be considered as a scoring system that is based on probabilistic models of linear sequences.

HMM profiles are able to capture the sequence properties for the set of peptide sequences. The profile generated can then be used to search for other sequences in a database that match this profile or a new sequence can be queried against the profile to see whether it matches the profile. This is possible as HMM inherently has probabilities assigned to each position of the alignment. In a typical HMM profile, the probability parameters are converted to additive log-odds scores before aligning and scoring a query sequence (Barrett *et al.*, 1997). Therefore, if the probability of the match state emitting residue y is p_y, and the expected background frequency of residue z in the sequence database is f_z, the score of residue z at this match state is log p_y/f_z. The scoring takes into account gap alignments also which is different from gap alignments in other sequence alignment tools like BLAST. In HMM profile, for an insertion of length x, there is a state transition into insert state which costs log $t_{MI.}$ t_{MI} is the state transition probability for

moving from match state to insert state, (x-1) state transitions for each subsequent insert state that cost log t_{II}, and a state transition for leaving the insert state that costs log t_{IM} (Eddy, 1998).

4.2.3 ANTIMIC Profile tool

The ANTIMIC profile tool integrated with the ANTIMIC database is based on HMMER (Eddy, 1998) (a program which uses Hidden Markov Models for motif description). The ANTIMIC profile tool is aimed at facilitating tentative classification of query sequences into different antimicrobial families. It uses predefined antimicrobial-specific library of profiles, and also allows users to generate profiles out of their specific sequences. The profile library has been created out of mature peptide regions of AMPs of different families as discussed below. The ANTIMIC profile tool suggests positions, which represent the signature for the selected family and potentially may be crucial for antimicrobial activity, as well as those, which are 'non-critical' in the functional domain of a family of sequences. The profiles used by this module can serve as templates for suggesting to which family of antimicrobial sequence a query sequence may belong. The use of profiles enables capturing of homologs from public databases, which have a high likelihood of belonging to a particular family.

The ANTIMIC profile tool has multiple modules . It consists of a profile-building module known as 'Build profiles' that enables the creation of profiles out of the sequences submitted by the user. The input sequences in this module can be in any format that is accepted by the program readseq,

(http://iubio.bio.indiana.edu/soft/molbio/readseq/). The module generates a Clustal W alignment of the sequences, which is used to generate the profile. The user can view the Clustal W alignment in the web browser. The result page gives the user the option to view the profile that has been generated or use the profile for querying. If the option of use profile is selected the user is directed to the 'Query profile' module. Using this module the user can input query sequences for query against the profile. The 'Query profile' module stores the profiles built by the user with an ID tag and stores a permanent profile library "antimicrobial.hmm". The antimicrobial.hmm consists of HMM profiles of several families of AMPs. The families currently included are melittin, magainin, bacteriocin, cecropin, and protegrin. HMM profiles of individual families are also provided separately.

The 'Query profile' module helps a user to predict to which family a query sequence most likely belongs to (based on primary sequence properties) and whether it is likely to share the same mode of action as the matched family of sequences. The results contain three sections: a ranked list of the best scoring HMMs; a list of the best scoring domains in order of their occurrence in the sequence; and alignments for the highest scoring domains. The matches are shown with scores (bits) and E-values. The bits score indicates how well the sequences match an HMM profile. E-value, which is calculated from bits score, shows the number of false positives that is expected to be seen at or above this bit score. Therefore, an E-value of 0.1 indicates that there is only a 10% space chance that the hit is a false or has come up by chance. Hence, a low E-value is best. The best hits appear on the top of the results list. The critical residues (highly conserved

residues) for both the query sequence and the consensus pattern for a family are shown in capital letters.

The second module is known as 'Query db'. Query db allows users to search for sequences in the GenBank 'nr' and ANTIMIC databases, which match specific profiles. These AMP profiles are predefined (for five AMP families) and could be used either as single profiles or as a library. Additionally, users may employ their own generated profiles.

4.3 HMM profiles of some AMP families

This section highlights a detailed analysis of HMM profiles generated for two AMP families (melittin, and beta-defensin) by ANTIMIC profile modules and its use in differentiating different query sequences.

4.3.4 Melittin profile analysis

Melittin are found in bees and are linear peptides without any disulfide bridges. They possess a highly asymmetric polar/non-polar amino acid distribution with six polar amino acids clustering at the c-terminal end (Maget-Dana, 1999). The peptides usually have a charge of +5 at pH 7 and a polar/non-polar amino acid ratio of 0.86 (Maget-Dana, 1999). Melittin is known to have a strong lytic activity towards red blood cells which is due to its amino acid residue tryptophan that plays a significant role in causing this hemolytic property (Blondelle *et al.*, 1993).

Six melittin sequences were taken and their mature peptide region extracted (**Table 4.3**). Using build profile module the sequences were aligned in a multiple sequence alignment and the HMM model was generated. A set of query sequences from

different sources were collected to test against the melittin HMM profile. This test dataset consisted of six analogs of melittin, which were different from the wild type melittin sequence by a few residues. All of these analogs had a substitution of an amino acid residue at different positions with tryptophan (W) residue (Blondelle *et al.*, 1993). Studies have shown that Trp residue plays a critical role in binding peptides to cholesterol present in biological membranes through the indole moiety (de Kruijff, 1990). It also plays role in hemolytic activity of thiol-activated sequences (de Kruijff, 1990). Hence, these analogs have been synthesized to understand the effect of a second Trp residue on melittin's hemolytic activity (Blondelle *et al.*, 1993). Two cecropin-melittin hybrid sequences were included which have a part of cecropin AMP sequence and a part of melittin AMP sequence. These hybrid sequences have been created in experimental labs studying the effect of hybridizing two different AMP sequences to get a more efficacious AMP sequence (Wade *et al.*, 1992, Juvvadi *et al.*, 1999). Protegrin AMP sequence from pig (PG3_PIG P32196) was included in the dataset. Non-AMP sequence Acyl-CoA dehydrogenase family member 8 (ACAD8_HUMAN) was introduced in the dataset. Finally, two melittin sequences were put in the dataset. One melittin sequence consisted of only the mature peptide region while the other was the complete peptide sequence, proprepeptide (**Table 4.4**).

The results of the query are shown in (**Appendix 1, Supplementary Figure 4.1**). The wild type, melittin mature peptide sequence and the melittin proprepeptide sequence both had the same E-value and the lowest E-value scores as expected. Hence, they were the closest sequences to the melittin profile. Protegrin and Acyl-CoA dehydrogenase family member 8 sequences were used as negative data to check against the melittin

profile. Both these sequences showed high E-value scores, which indicated their distance from the melittin profile. Acyl- CoA dehydrogenase had slightly lower score than protegrin as it was a longer sequence length and thus had more number of random matches.

Next were the six melittin analogs all of which had nearly similar E-value scores. Mut5_L6 and mut13_L13 both had the same E-value scores, which was the lowest score in the group of melittin analogs. These two analogs have a substitution of Leu-> Tryptophan at position 6 and 13 respectively. Experimental evidence shows that a substitution of any of these two leucines with Trp leads to a decrease in hemolytic activity. The mut13_P14 is different from other analogs due to a Proline (P->W) (Tryptophan) substitution. This substitution leads to a small increase in hemolytic activity (Blondelle *et al.*, 1993). Not surprisingly, this sequence had a slightly greater E-value than the rest of the analogs. The analogs were followed by the cecropin-melitin hybrid sequence (CecropinA(1-8)-Mel(1-18)). This was closer to the melittin profile as it had a greater part of melittin sequence in its sequence length. CA (1-7)M(2-9) on the other hand had a shorter melittin sequence contributing to the hybrid formation and hence did not show a favorable E-value score.

This result shows that HMM based scoring system can be used to segregate sequences having different properties into groups based on differences in E-values. Analysis of the E-values of different test sequences shows that the melittin profile generated by HMM is able to differentiate between members of the melittin family and non-members. It can also differentiate sequences where residues are substituted at critical position that directly affects the function of the melittin sequence. Hence, this tool can be

used to create specific profiles out of analogs or mutated sequences to test against new query sequences for checking profile similarity. The next example demonstrates this point.

4.3.4 Melittin analog profile analysis

Another analysis that was done using HMM was to classify analogs of a particular family of AMP and create profiles out of it to observe the critical residues that cause certain properties of the AMP to increase or decrease. Melittin analogs were collected from literature (Blondelle *et al.*, 1993) that show change in the hemolytic activity of melittin. Two different profiles were generated from analogs. One profile was created out of three analog sequences that had substitution of leucine residues with tryptophan at position 9, position 13 and position 16 of the wild type melittin residue. These sequences were observed to show decreased hemolytic activity in assays (Blondelle *et al.*, 1993). A second profile was created out of seven analog sequences with tryptophan substitutions at positions 1, 7, 11, 12, 15, 23, 21. These sequences showed significant increase in the hemolytic activity of melittin.

Next, these profiles were tested against a set of sequences for their specificity to see if they could differentiate between a sequence that has increased hemolytic activity and one that has decreased hemolytic activity. A set of four sequences was chosen to test against the profiles. Two were analogs with substitution of lysine-23 with tryptophan and leucine-16 with tryptophan. Lysine-23 substituted analog shows increased hemolytic activity while leucine-16 has decreased hemolytic activity (Blondelle *et al.*, 1993). Mel_apicc mature peptide sequence representing the wild type melittin sequence was the third sequence in the test set. The fourth sequence was an analog with isoleucine

substituted with tryptophan at position 2. This sequence did not show any significant increase or decrease in activity compared to the wild type melittin sequence.

Querying against the "increase hemolytic activity" profile showed that the wild type melittin sequence (mel_apicc) and K-23 (increased hemolytic activity) analog had the lowest E-values*. Leucine-16 that represents the analog with decreased hemolytic activity had a higher E-value than K-23. I-2 analog that showed no change in hemolytic activity and had a slightly higher E-value than L-16. Thus, the profile was able to differentiate the analog with increased activity from the one with decreased activity.

It was observed that querying against the "decrease hemolytic activity" profile using the same test set, leucine-16 (L-16) and wild type melittin sequence had lower E-values than K-23 and I-2. Thus, L-16 was closer to this profile, an obvious outcome and K-23 was more distant to this profile as compared to L-16.

Since the analogs have single substitutions in their sequences, the E-values to differentiate the two different categories of melittin analogs were not on a very wide scale difference. However, a significant difference in the E-values was observed that enabled ranking them on the basis of closeness to the profile. The test set sequences and the results of the analog profiles are in (**Table 4.5, and Appendix 1, Supplementary Figure 4.2**).

*Low E-values indicate the query is closer to a given a HMM profile

4.3.4 Beta-defensin profile analysis

The melittin profile is an example of model dataset that is fairly conserved and homogenous in its sequences. Therefore, another AMP family was taken where the peptide sequences are not so well conserved among themselves, though they have been classified under the same beta-defensin family. The beta-defensin profile was made of 13 different beta-defensin mature peptide sequences from human, mouse, and different monkey species (**Table 4.3**). The test dataset contained five sequences, which consisted of two beta defensin from different monkey species and one beta defensin from goat. Acyl-CoA dehydrogenase family member 8 (non-AMP) and protegrin (**Table 4.4**). **Supplementary Figure 4.3** shows the results of the querying against the beta-defensin profile. Beta-defensin from BD01_CERPR (Preuss' monkey) and BD01_PONPY (Orangutan) had the most favorable E-values (low E-value). Beta-defensin from goat had a different E-value indicating that it was not very close to the primate and rodent beta-defensin sequences. The goat beta defensin sequence has residue substitutions in many conserved positions. The cysteine residue positions are conserved. However, its E-value is much lower compared to non beta-defensin sequence protegrin and non-AMP sequence (Acyl-CoA dehydrogenase family member 8) and is comparable to the other two monkey beta-defensin sequences. Protegrin and Acyl-CoA dehydrogenase family member 8 had very high E-value scores indicating they did not belong to this AMP profile.

4.3.4 Querydb results

Querydb enables to extract sequences from public databases that have similar sequence properties. NR (non-redundant) peptide database of NCBI is the public database , that has

been integrated to the ANTIMIC profile module. The nr dataset contains 137,010 peptide sequences.

The melittin profile was searched against the nr database with a default E-value cutoff of 10 to test its specificity and sensitivity. The search returned three hits (**Appendix 1**, **Supplementary Figure 4.4**). The first two hits (gi|69550, gi|229444) with low E-value scores belonged to the melittin family. The third hit (gi|16121500) was a tyrosine-specific transport protein from the bacteria *Yersinia pestis CO92* which does not belong to the melittin family. It perhaps appeared since some conserved residues of melittin domain matched the residues of the tyrosine-specific transport protein. This profile has a high sensitivity and fair specificity index and a good correlation coefficient (**Table 4.6**).

The beta-defensin profile was searched against nr dataset with an E-value cutoff of 10. Search returned 12 hits (**Supplementary Figure 4.5**). The top most hit was beta-defensin 1 from human which was also one of the sequences of the dataset used in creating the beta-defensin HMM profile. Majority of the hits were beta-defensins from different mammalian species (human, bovine, mouse, horse). Though the profile was generated using a number of primate species sequences, the primate sequences did not come up as hits since the nr dataset used for querying lacked monkey beta-defensins. The only sequence that was a false hit was, gi|230338 which is a trypsin peptide complex with Bowman-birk inhibitor. The sensitivity, specificity and correlation coefficient for beta-defensin profile indicates that it has average sensitivity, high specificity and a fairly good correlation coefficient. **Table 4.6** gives the sensitivity, specificity and correlation coefficient of the results from querying melittin and beta-defensin profile against nr

database. The overall quality of the profile search against the nr database has been calculated in terms of sensitivity, specificity and correlation coefficient with the following formula:

Sensitivity (Sn) = TP/ (TP+FN) TP =True Positive; FN=: False negative; FP=: false positive

Specificity (Sp) = TP/ (TP+ FP)

Correlation coefficient (CC) = (TP*TN)- (FN*FP)/√(TP+FN)(TN+FP)(TP+FP)(TN+FN)

4.4 Discussion

HMM has been used as the method for creating peptide profiles of different AMP profiles. The author has taken examples of different families of AMPs and has attempted to show that using HMM profiles one can predict the salient functional residues and a possible change in the strength of the property, even with single mutations at some residues. This has been shown through the melittin analog properties example. This example can be extrapolated to design *in-silico* mutant peptides, which have a desired property provided *apriori* knowledge about a family of sequences exists. It has also been possible to differentiate between sequences that are evolutionary divergent though they belong to the same AMP family. This has been shown with the beta-defensin profile analysis.

As a part of future work, HMM profiles can be created for the AMP families that have not been covered by this study. Comparison of HMM with other machine learning methods like ANN, SVM was beyond the scope of this thesis. These methods can be

tested on AMP families to compare performance of HMM with other methods. These methods can also be combined with HMM to see if a better classification method can be determined.

4.5 Conclusion

This study was to investigate the classes of AMP peptides and see if current classifications fit, and attempt to propose a computational method of classification that could be used across all AMPs based on the sequence properties. The HMM profiles were also set up find new AMPs. However, as previously reported and also observed, the variation and diversity of the AMP sequences even within the same family and species (Maxwell *et al.*, 2003) makes is difficult to identify or predict new AMPs. Thus, a new approach is proposed that has been used recently for some specific classes of genes (Frech *et al.*, 1997, Wasserman and Fickett, 1998) based on the model of the gene's promoter region. This approach seems reasonable to use for the purpose of AMP gene discovery as literature reviews suggest that the promoter regions of the highly diverse AMPs are fairly conserved (Ganz, 2003). This approach can be suitably complemented with homology based gene identification methods to increase the possibilities of extracting new AMPs from whole genomes. Chapter 5 and onwards shows implementation of this approach.

Table 4.3: Sequences from melittin and beta-defensin AMP family used to create HMM profiles

melittin	peptide_name	Species	sequence (mature_peptide)
	MEL_APICC	Apis cerana cerana	GIGAVLKVLTTGLPALISWIKRKRQQ
	MEL_APICE	Apis Cerana	GIGAVLKVLTTGLPALISWIKRKRQQ
	MEL_APIDO	Apis dorsata	GIGAILKVLSTGLPALISWIKRKRQE
	MEL_APIFL	Apis florae	GIGAILKVLATGLPTLISWIKNKRKQ
	MEL_VESMC	Vespula maculifrons	GIGAVLKVLTTGLPALISWIKRKRQQ
	MEL_APIME	Apis mellifera	GIGAVLKVLATGLPALISWIKRKRQQ
beta-defensin			
	BD01_MOUSE	Mus musculus	DQYKCLQHGGFCLRSSCPSNTKLQGTCKPDKPNCCKS
	BD01_HUMAN	Homo sapiens (Human)	DHYNCVSSGGQCLYSACPIFTKIQGTCYRGKAKCCK
	BD01_PRECR	Presbytis cristata (Silvered langur)	DHYNCVRSGGQCLYSACPIYTKIQGTCYHGKAKCCK
	BD01_PREME	Presbytis melalophos (Banded langur)	DHYNCVRSGGQCLYSACPIYTKIQGTCYHGKAKCCK
	BD01_PREOB	Presbytis obscurus (Dusky langur)	DHYNCVRSGGQCLYSACPIYTKIQGTCYHGKAKCCK
	BD01_HYLLA	Hylobates lar (Common gibbon)	SDHYNCVRSGGQCLYSACPIYTKIQGTCYQGKAKCCK
	BD01_GORGO	Gorilla gorilla gorilla	DHYNCVSSGGQCLYSACPIFTKIQGTCYGGKAKCCK
	BD01_MACFA	Macaca fascicularis (Crab eating macaque)	DHYNCVRSGGQCLYSACPIYTRIQGTCYHGKAKCCK
	BD01_MACMU	Macaca mulatta (Rhesus macaque)	DHYNCVRSGGQCLYSACPIYTRIQGTCYHGKAKCCK
	BD01_CERAE	Cercopithecus aethiops (Green monkey)	DHYNCVRSGGQCLYSACPIYTKIQGTCYHGKAKCCK
	BD01_CERER	Cercopithecus erythrogaster (Red-bellied monkey)	HYICVRSGGQCLYSACPIYTKIQGTCYHGKAKCCK
	BD01_PAPAN	Papio anubis (Olive baboon)	DHYNCVRSGGQCLYSACPIYTRIQGTCYHGKAKCCK
	BD01_CERPR	Cercopithecus preussi (Preuss's monkey)	DHYNCVRSGGQCLYSACPIYTKIQGTCYHGKAKCCK

Table 4.4: Sequences queried against melittin and beta-defensin profiles

Name	Sequence
melittin test sequences	
mel_apicc (mature_peptide) (mellitin1) wildtype	gigavlkvlttglpaliswikrkrqqg
mut5_l6 (mutant mellitin)	gigavwkvlttglpaliswikrkrqq
mut13_l13 (mutant mellitin)	gigavlkvlttgwpaliswikrkrqq
mut1_g1 (mutant mellitin)	wigavlkvlttglpaliswikrkrqq
mut6_l7 (mutant mellitin)	gigavlwvlttglpaliswikrkrqq
mut10_t11 (mutant mellitin)	gigavlkvltwglpaliswikrkrqq
mut13_p14 (mutant mellitin)	gigavlkvlttglwaliswikrkrqq
cecropina(1-8)-melittin(1-18) (mellitin hybrid)	kwklpkkigigavlkvlttglpalis
ca(1-7)m(2-9)	kwklfkkigavlkvl
Protegrin (PG3_PIG)	glcycrrrfcvcv
acyl-coadehydrogenasefamilymember8 (ACAD8_HUMAN)	mlwsgcrrfgarlgclpgglrvlvqtghrsltscidpsmglneeqkefqkvafdfaaremapnmaewdqkelfpvdvmrkaaqlgfggv yiqtdvggsglsrldtsvifealatgctsttayisihnmcawmidsfgneeqrhkfcpplctmekfasycltepgsgsdaaslltsakkqgdh yilngskafisgagesdiyvvmcrtggpgpkgiscivvekgtpglsfgkkekkvgwnsqptravifedcavpvanrigsegqgfliavrglng griniascslgaahasviltrdhlnvrkqfgeplasnqylqftladmatrlvaarlmvrnaavalqeerkdavalcsmaklfatdecfaicnqal qmhggygylkdyavqqyvrdsrvhqilegsnevmrilisrsllqe
mel_apicc(complete peptide) (melittin_complete)	mkflvnvalvfmvvyisfiyaapepepapeaeaeadaeadpeagigavlkvlttglpaliswikrkrqqg
beta-defensin test sequences	
acyl-coadehydrogenasefamilymember8	mlwsgcrrfgarlgclpgglrvlvqtghrsltscidpsmglneeqkefqkvafdfaaremapnmaewdqkelfpvdvmrkaaqlgfggvyiqtdvggsglsr ldtsvifealatgctsttayisihnmcawmidsfgneeqrhkfcpplctmekfasycltepgsgsdaaslltsakkqgdhyilngskafisgagesdiyv vmcrtggpgpkgiscivvekgtpglsfgkkekkvgwnsqptravifedcavpvanrigsegqgfliavrglnggriniascslgaahasviltrd hlnvrkqfgeplasnqylqftladmatrlvaarlmvrnaavalqeerkdavalcsmaklfatdecfaicnqalqmhggygylkdyavqqyvrd srvhqilegsnevmrilisrsllqe
Protegrin	gglcycrrrfcvcv
bd01_cerpr	dhyncvrsggqclysacpiytkiqgtcyhgkakcck
bd01_caphi	qgirsrrschrnkgvcaltrcprnmrqigtcfgppvkccrkk
bd01_ponpy	sdhyncvssggqclysacpiftkiqgtcyrgkakcck

Table 4.5: Sequences queried against melittin analog profiles

Name	sequence
melittin analogs	
K-23	GIGAVLKVLTTGLPALISWIKR*W*RQQ
L-16	GIGAVLKVLTTGLPA*W*ISWIKRKRQQ
I-2	G*W*GAVLKVLTTGLPALISWIKRKRQQ
melittin wild type	
mel_apicc (mature_peptide)	GIGAVLKVLTTGLPALISWIKRKRQQG

Table 4.6: Sensitivity, Specificity, Correlation coefficient calculation

AMP profile	TP	FN	FP	TN	Sn	Sp	CC
melittin	2	0	1	137008	1	0.66	0.99
beta-defensin	11	9	1	136990	0.55	0.91	0.71

TP: true positive; FN: false negative; FP: false positive; TN: true negative;
Sn: sensitivity; Sp: Specificity; CC: correlation coefficient

Part III:Chapter 5: *Ab-initio* search for TFBS motifs

Nothing great was ever achieved without enthusiasm.
(Ralph Waldo Emerson)

5.1 Introduction

From the previous chapter (Chapter 4), analysis of AMP sequences within a single family like defensins showed that there is considerable amount of variation in sequences even within the same family. This attributes to their ability to have a broad spectrum of antimicrobial activity (Pereira, 2006). Due to their inherent variability in sequence, AMPs demonstrate low levels of similarities for homology to be inferred (Patil *et al.*, 2004, Maxwell *et al.*, 2003, Hughes, 1999) thus, one alternative is to look at the regulatory regions of these AMPcgs to see if they are more homologous in terms of the regulatory elements.

The author examined the regulatory regions of AMP genes in the effort to investigate the presence of conserved motifs upstream of the highly diverse AMPs gathered in the ANTIMIC database described in Chapter 3. In particular, transcription factor binding site (TFBS) motifs were closely investigated. One of the main features of commonality amongst the highly diverse AMPs, and across AMP families, is their involvement in some kind of defense or defense related responses. It is therefore possible in principle that common regulatory mechanisms are involved in triggering their expression in response to an external threat. Their expression may be regulated by common transcription factors (TFs) that regulate the expression at the transcriptional level. Hence, the aim in this chapter was to uncover TFs or TF groups that are common to AMP genes, whose presence could be put into biologically relevant contexts for transcriptional regulation of AMP genes. Unsurprisingly, motifs discovered by *ab-initio*

methods show common features across various AMP gene families and some appeared to be specific to certain AMP families.

Studies have been done using computational and experimental approaches to find conserved TFBS motifs across the same AMP gene within various species. For example, lactoferrin is a serum tranferring protein that is involved in the transport of ions (Fe3+) and in human and bovine is known to have antimicrobial activity (Bellamy *et al.*, 1992, Bellamy *et al.*, 1993). An analysis done on the promoter region of this gene from multiple species (human, mouse, bovine and porcine species) showed that they had some conserved regulatory elements. A non-canonical TATA box (GATAAA) with an adjacent Sp1 site was present in all the promoter regions. All the promoters had similar basic arrangement and a GC-rich sequence. Moreover, in two species, human and mouse, multiple steroid hormonal response elements specific only to these two species were found (Teng, 2002). However, there has been no attempt so far to find common motifs across different AMP genes across different species. This study demonstrates the first attempt to find common and taxon-specific motifs in a large scale manner across the AMP families based on the databases described in Chapter 3.

This study has been the first in attempting to find common and specific motifs in such a large-scale manner across many AMP families.

5.2 Background

5.2.1 Basic introduction of transcription, the key process involved in gene regulation

Transcription is a complex process of decoding information present in DNA into mRNA molecules. This process depends on the collective action of transcription factors along with the core RNA polymerase II transcriptional machinery, and a variety of co-regulators that bridge the DNA binding factors to the transcriptional machinery. In addition a number of chromatin remodeling factors that mobilize nucleosomes, and an array of enzymes that catalyze the covalent modification like acetylation, decacetylation, phosphorylation, metyhylation etc of histones and other proteins are also required (Kadonaga, 2004)

Initiation of transcription requires the enzyme RNA polymerase and transcription factors. Transcription factors initiate transcription, but are not themselves part of RNA polymerase. The focus will be on RNA polymerase II and its promoter region as it is responsible for mRNA transcription. Polymerase II is not capable of initiating transcription on its own, without the co-factors. This is to check against unscheduled transcription, which can be disastrous for a cell.

There are two major steps in the initiation of transcription. The first step is binding of different transcription factors (TFs) to upstream promoter and enhancer sequences to form a multi-protein complex. In the second step, this complex directly or indirectly recruits a polymerase II complexed with some general transcription factors (GTFs) to the core promoter. Subsequently, transcription is initiated by this initiation complex, which itself is subject to regulatory influences of TFs.

Transcriptional initiation is activated by two types of cofactors. They are transcriptional accessory factors (TAFs) and GTFs like TFIID, TFIIE, TFIIH, and TFIIF. The TAFs form the TFIID complex. TFIID binds to TATA box via TATA box binding protein (TBP). TFIID is involved in the transcription of most pol II promoters. TFIIE and TFIIH are two GTFs that are necessary for pol II to clear the promoter for elongation. TFIIF is required for bringing pol II into closer contact with the promoter region during the initiation process. In addition to these GTFs, there are several other transcriptional activators and repressor proteins (TFs) involved in transcriptional regulation. Only specific subsets of these factors bind directly with TAFs or form a ternary complex with TAF. Once the complete complex including TFs, TAFs, GTFs and pol II is assembled on the promoter, this is called the initiation complex, which is now competent to initiate RNA synthesis.

5.2.2 Defining a eukaryotic promoter

A eukaryotic promoter is defined as the region containing binding sites for transcription factors. RNA polymerase itself binds around the start point of transcription initiation on the gene, but does not directly contact the extended upstream region of the promoter. The difference between eukaryotic and prokaryotic promoter is that initiation at eukaryotic promoter involves a large number of factors that bind to a different *cis*-acting element. Bacterial promoters are largely defined in terms of the binding site for RNA polymerase in the immediate vicinity of the start point. The promoter region for RNA polymerase II is usually upstream of the start point of a gene beginning from the start of the first exon. Each promoter consists of characteristic sets of short conserved sequences that are recognized by appropriate class of factors. These *cis*-acting sites are usually spread over a

region of >200bp. Some of the elements and the factors that recognize them are common; they are found in a variety of promoters and are used constitutively. Others are specific, they identify particular class of genes and their use is regulated. The elements occur in different combinations in individual promoters. All RNA polymerase II promoters have sequence elements close to the start point that are bound by the basal apparatus and that establish the site of initiation. Sequences positioned further upstream determine whether the promoter is expressed in all cell types or is specifically regulated. Promoters that are constitutively expressed have upstream sequence elements that are recognized by ubiquitous activators. Promoters that are expressed only in certain times or places have sequence elements that require activators that are available only at certain times or places. Structurally, promoters contain the transcription start site (TSS) and contain a part of the first exon of a gene.

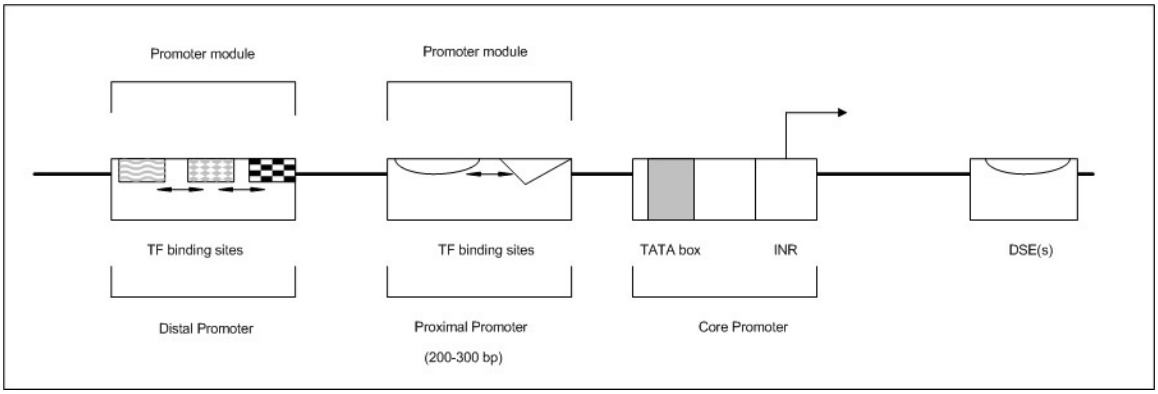
A RNA polymerase II eukaryotic promoter contains different types of promoter elements in its structure. They are core promoters, proximal promoters, distal promoters, enhancer, silencers, boundary /insulators. (Butler and Kadonaga, 2002)

Core promoters are usually within -35 to +35 region of promoter and contain the transcription start site (TSS). They constitute the general transcription factor binding sites involved in initiation of transcription like TATA box, Inr (initiator), BRE (TFIIB recognition element),DPE (downstream core promoter element). Each of these motifs have a specific function in the process of transcriptional regulation. It is important to note that each of these core promoter elements is found in some but not all core promoters. For example, TATA box is not found in all core promoters. In addition to the core promoter, other *cis*-acting DNA sequences that regulate RNA polymerase II transcription

include the proximal promoter, enhancers, silencers, and boundary/insulator elements. These elements contain recognition sites for a variety of sequence-specific DNA-binding factors that are involved in transcriptional regulation. The proximal promoter is the region in the immediate vicinity of the minimum promoter site (roughly from −250 to +250 nt). The minimum promoter is the region that is capable of initiating basal transcription and may include a few more sites located close to the TATA box or the TSS. The proximal promoter contains the functionally important regulatory controls and is present near the TSS. The distal part of promoter is also the most variable one with respect to composition as well as length. It can consist of binding sites for any of the transcription factors.

Enhancers and silencers can be located many kilo base pairs from the transcription start site and act either to activate or to repress transcription. Boundary/insulator elements appear to prevent the spreading of the activating effects of enhancers or the repressive effects of silencers or heterochromatin (Butler and Kadonaga, 2002). **Figure 5.1** shows a graphical representation of the various promoter regions on the genome.

Figure 5.1: Schematic diagram of the different regions of a polymerase II promoter
The shaded boxes, semicircle and triangle indicate the TF binding sites.DSE: Distal Sequence Element, INR:Initiator This diagram has been redrawn from (Werner, 1999).



5.2.4 Computational methods for identification of regulatory elements in promoter region

Many computational methods for predicting promoters have been developed over the last few years. In general, the algorithms can be divided into two groups. First, is the signal-based approach, which relies on the recognition of relatively conserved signals and conserved spacing among patterns such as the TATA box, CCAAT box. Second, there is the content-based approach, which distinguishes promoter sequences from non-promoter sequences based on content differences such as triplet base-pair preferences around the TSS, hexamer frequencies in conservative 100-bp upstream regions, etc. using linear discriminant function (TSSG, TSSW) (Werner, 1999) or quadratic discriminant analysis (CorePromoter) (Werner, 1999). These programs have been able to predict about 13%–54% of the promoters, correctly; each program also predicted a number of false positive promoters. To find the proximal promoter, the approach is to find the TSS.

However specification of the TSS can be difficult. It is also known that a growing number of genes have more than one TSS close to each other, known as alternative start

sites. Many algorithms that do promoter prediction are based on EPD (Eukaryotic Promoter Database). This database contains experimentally elucidated promoter regions for many eukaryotic species. Detection of exact location of TSSs is not a trivial problem and is often confronted with issues of false predictions. Algorithms that detected TSSs were based on the identification of TATA box sequences, which are often located ~30bp upstream of a TSS. However, TATA binding motif is found very frequently in the upstream region as much as in every 250 bp in long genome sequences, reflecting the promiscuous binding characteristics of the TATA-like sequences and thus this does not prove to be an effective approach. Newer algorithms have shifted the emphasis to the prediction of promoters that contain one or more TSS(s). This approach is biochemically more justified as many genes have multiple TSS(s).

In human genome, the sequence property that is used to predict promoter region is based on differences in methylation of CpG dinucleotides. There are regions in the genome sequences >200 base pairs that have high G+C content, and are known as CpG islands. CpGs are methylated on cytosine as a phenomenon for regulation of gene activity. However, in regulatory sequences, like promoter region CpGs remain unmethylated unlike other regions where the CpG methylation can be up to 80%. (Wasserman and Sandelin, 2004) Methylated cytosines are mutated to adenosines at a high rate, resulting in a 20% reduction of CpG frequency in sequences without a regulatory function as compared with the statistically predicted CpG concentration. This imbalance in CG dinucleotide has been exploited in bioinformatics for detecting promoter sequences. Numerous methods have been developed that directly or indirectly detect promoters on the basis of the CG dinucleotide imbalance. The simple methods

based on frequency of CpG dinucleotides perform remarkably well at correctly predicting regions that are proximal to or that contain the sites of transcription initiation. Two commonly known methods – Eponine and FirstEF use divergent approaches. FirstEF finds regions in genes with higher concentration of CG dinucleotides than the local C and G concentrations would suggest. It subtly improves performance by restricting predictions to those regions that contain or are followed by a predicted 3'- splice site, thereby indicating the presence of a first exon. Eponine uses a neural network model that analyses the over-and under-represention of longer oligonucleotide sequences. As Eponine's strand prediction is based on the identification of a TSS, which is an unreliable step, predictions of promoter orientation are not reliable. There is also the phenomenon of the presence of bidirectional promoters, which limits the ability of the current bioinformatics methods to accurately predict promoter orientation.

It would be worthwhile to point to two recent programs for finding promoter regions- Dragon Promoter Finder (DPF) (Bajic *et al.*, 2002) and Dragon Gene Start Finder (DGSF) (Bajic and Seah, 2003). DPF does a content analysis of the region around the predicted. It uses aritificial neural network (ANN). DGSF also uses ANN along with CpG islands and DPF output (Bajic *et al.*, 2004). In a recent study on the whole human genome, DGSF appeared to be the most accurate promoter prediction program, while DPF was one with the second highest sensitivity.

Not all transcription initiation sites are proximal to CpG islands and that the association between CpG dinucleotides and promoters is not present in all organisms. As only ~60% of human promoters are situated proximally to CpG islands, hence alternative approaches are required to identify a substantial portion of promoters. The identification

of promoter regions that lack CpG islands requires the use of transcript data. Recurrent alignment of the 5'edges of ESTs and /or full-length cDNAs can be indicative of promoter locations. Two programs that are based on mapping the 5' most position (5' untranslated mRNA sequence) of full-length cDNA to genome are – PromoSer (http://biowulf.bu.edu/zlab/PromoSer/) and FIE2 (http://research.i2r.a-star.edu.sg/FIE2.0/). PromoSer identifies the TSS of a gene, by mapping all available mRNA and EST sequence data onto the genome and then tracks the overlapping alignments (denoted as a *cluster*) to determine the furthest possible extension to these sequences and hence determines the TSS. In many cases, PromoSer data set is enriched with full-length mRNA sequences produced by cap-trapping and oligo-capping methods, that facilitates higher confidence in the predictions. **Table 5.1(a, b)** list the promoter databases and the prediction tools.

Table 5.1a: Promoter databases

Promoter databases		
Source	URL Address	Method of extraction (Data quality)
Genomatix (GPD)	http://www.genomatix.de/	Experimental verfied TSS (gold standard)
TRED	http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home	1.known,curated (collected from EPD, DBTSS, GenBank) ; 2.predicted
DBTSS	http://dbtss.hgc.jp/	oligo-capped cDNAs (experimentally confirmed full length cDNA) mapped to genome,alternate TSS accountable
BU (PromoSer)	http://biowulf.bu.edu/zlab/PromoSer/	map mRNA+EST-genome (full length mRNA data from refseq, oligo-captrapping in some cases. Alternative TSS sites accountable
UCSC	http://genome.ucsc.edu/	This includes only cases where the transcription start is annotated separately from the coding region start. Sequences 5000 bases upstream of annotated transcription start of RefSeq genes.
EPD	http://www.epd.isb-sib.ch/	experimental
Ensembl	http://www.ensembl.org/index.html	pulls out upstream region based on EMBL mRNA records. No guarantee that the upstream regions are promoter regions and the TSS is right.
Mpromdb	http://bioinformatics.med.ohio-state.edu/MPromDb/	Based on experimentally found TSS
H-invitational database (Only TSS info)	http://www.jbirc.aist.go.jp/hinv/index.jsp	Experimental (full length cDNA)
Riken (Only TSS info)	http://fantom.gsc.riken.go.jp/	CAGE tags (experimental)

Table 5.1b: Promoter prediction tools

Promoter prediction programs	URL Address
Eponine	http://www.sanger.ac.uk/Software/analysis/eponine
FirstEF	http://rulai.cshl.edu/tools/FirstEF
Promoter Scan	http://biosci.umn.edu/software/proscan/promoterscan.htm
TSSG/TSSW	http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html
FunSiteP	http://transfac.gbf.de/dbsearch/funsitep/fsp.html
NNPP	http://www-hgc.lbl.gov/projects/promoter.html
PromFD	http://beagle.colorado.edu/~chenq/Hypertexts/PromFD.html
Dragon Promoter Finder	http://research.i2r.a-star.edu.sg/promoter/promoter1_5/DPF.htm
Dragon Gene Start Finder	http://research.i2r.a-star.edu.sg/promoter/dragonGSF1_0/genestart.htm
PromoterInspector	http://www.genomatix.de/

FIE2 (5' end Information Extraction v2) is another web based program that identifies and extracts nucleotide sequence region around the start of genes (promoter region) and their translation initiation site (TIS). It uses information provided by the National Center for Biotechnology Information's (NCBI's) LocusLink. FIE2 identifies the 5'-most end of a gene on its respective chromosome based on alignment of a selected set of mRNAs representative of the gene. The accuracy of the information extracted is therefore limited by the accuracy and completeness of the sequence annotation with regard to the completeness of the cDNA sequences till the 5' untranslated region and sequence alignment provided by Locus Link. In addition, multiple TIS positions are also occasionally presented, for example, as a result of multiple alignments of transcript variants.

The latest technique that brings us closer to accurate promoter prediction is CAGE tag transcripts. CAGE (Cap analysis of gene expression) is a cap-cloning technique that has been extended with a SAGE-like procedure to cleave the initial 5' 20 nucleotides of full-length cDNAs. These oligomers are then ligated into long polymers and sequenced. Generation of these CAGE tags from transcripts that are derived from diverse tissues promises not only to facilitate improved promoter prediction, but also to provide insights into tissue-specificity.

5.2.4 Detection of transcription factor binding sites

DNA sequences that are a part of the promoter region do not give direct information about regulation. Promoters do not have fixed stretches of sequence homology, which are responsible for promoter function. The elements influencing transcriptional regulation that binds to promoter regions do so in short stretches of the region. These regions or motifs are known as transcription factor binding sites (TFBSs). TFBSs are motifs that are usually very short (5-30 nucleotides) and

gapless. These sites are interspersed with non-conserved sequences. The regulatory regions (promoters) that contain regulatory sites are very long (varying from several hundred to more than 1000 nucleotides). The actual regulatory DNA sites corresponding to a motif are called the instances of that motif. Every instance of a motif normally has the same length, but they may have slightly different sequence compositions. This variability of regulatory sites makes biological sense. Better gene expression control can be achieved by having regulatory sites with different intrinsic affinities for regulatory proteins. TFBSs do not show any significant specific pattern with respect to location and orientation within the promoter sequences. Identification of TFBSs computationally poses a problem since they are very short signals and have sequence variability that is not very well understood (Tompa *et al.*, 2005).

There are mainly two different computational approaches to detect TFBSs namely finding motifs with known TFBS Position Weight Matrices (PWMs) and secondly *ab-initio* motif search. The first approach is finding TFBSs on a sequence using matrix or other models of known TFBS. The binding sites are determined by experimental methods like deletion mapping and then mutagenesis of the regulatory sequences (TFBSs). A single TF can bind multiple target sequences having significant variation; hence, multiple sites are required to construct a model. The multiple binding sites are aligned. The sequence variability of the collection of binding sites strongly affects the downstream models for predicting additional sites. A consensus sequence is generated from the alignments of the multiple binding sites. To accurately reflect the characteristics at each position, a matrix that contains the number of observed nucleotides at each position is created. This is known as the position frequency matrix.

The frequency matrix is converted to a position weight matrix (PWM) in which normalized frequency values are converted to log-scale. PWM are also known as PSSM (position-specific scoring matrices). Since TFBSs are short, degenerate sequence motifs that can occur very frequently across a whole genome, the PWM provides a summary of the binding specificity of these TFs and hence is a representation of their binding specificity. Using the PWM, a DNA sequence can be scanned for known TF binding-site elements. Several programs have been developed to perform searches based on PWM and IUPAC: SIGNAL SCAN, MATRIX SEARCH, MatInspector, ConsInspector, TFSearch, etc. PWM based search is considered sensitive, however there are a few drawbacks of using PWM. Using PWM approach will yield only a small fraction of the predicted binding sites, which are functionally significant.

Current matrix models are based on the assumption that a nucleotide at one position has no effect on the likelihood of a nucleotide being observed at an adjoining position. For a few cases in which large data collections have been generated to richly define binding, advanced models that incorporate higher-order interactions between positions have proved more effective (Wasserman and Sandelin, 2004).

Another assumption is that TFs have strict spatial requirements in their binding sites that preclude variable spacing (Wasserman and Sandelin, 2004). For some TFs, such as subset of the nuclear receptor family, variable spacing is allowed, rendering standard PWMs inappropriate for TFBS prediction.

Another limitation of the matrix model based TFBS prediction is the construction of models for predicting binding sites for TFs is limited by the limited number of valid *cis*-regulatory elements.

The recent advancement of microarray technology and the availability of a large number of complete genome sequences have resulted in a new approach to finding TFBSs. Genes are classified under different clusters based on their expression patterns. Genes in the same cluster are assumed to be co-regulated. However, it should be noted that co-expressed genes which are not co-regulated may not necessarily share same promoter features (Werner *et al.*, 2003).

Computational approaches like a*b-initio* TFBS detection method can be used to discover regulatory elements. In *ab-initio or de-novo* approach, for a given set of co-regulated genes or genes belonging to same family, programs detect over-represented motifs in the regulatory regions. Some of the programs that use this approach are listed in **Table 5.2**. A prior knowledge of TFBSs is not needed in this approach and hence it is more relevant for searching new and highly conserved motifs within promoter regions as well as getting already known TFBSs (van Helden, 2003). Dragon Motif Builder, the program used in this thesis will be discussed in detail in this chapter. Appendix2 lists the DMB parameters.

Table 5.2: Programs for *de novo* prediction TFBS motifs.

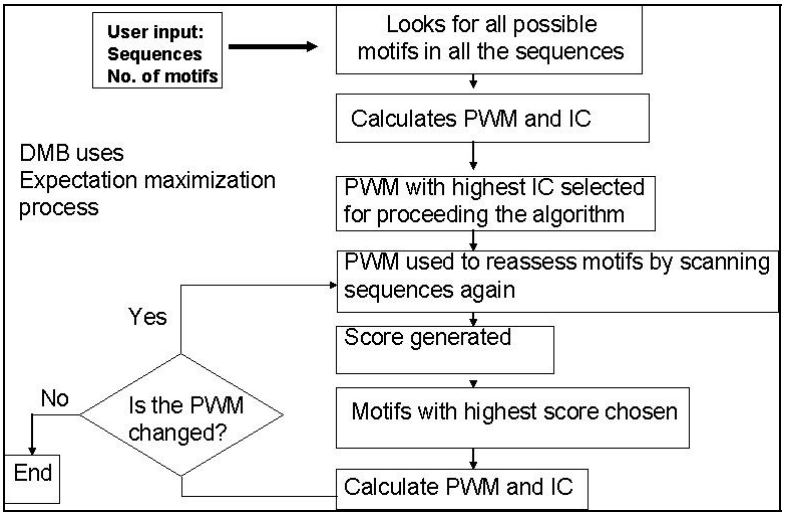
This table has been taken from Tompa *et al.*, 2005). ****PMID (Pubmed Unique Identifier)***

Program	**Operating Principle**	**URL Address**	**Reference (*PMID)**
Align ACE	Gibbs Sampling algorithm that returns a series of motifs as weight matrices that are over-represented in the input set	http://atlas.med.harvard.edu/	10698627
ANN-Spec	Models of the DNA-binding specificity of atranscription factor using a weight matrix	http://www.cbs.dtu.dk/~workman/ann-spec	10902194
Consensus	Model motifs using weight matrices, searching for the matrix with maximum information content	http://bifrost.wustl.edu/consensus	10487864
GLAM	Gibbs sampling-based algorithm that automatically optimizes the alignment width and evaluates the statistical significance of its output.	http://zlab.bu.edu/glam	14704356
Improbizer	Uses expectation maximization to determine weight matrices of DNA motifs that occur improbably often in the input sequences	http://www.soe.ucsc.edu/~kent/improbiser	15375261
MEME	Optimizes the E-value of a statistic related to the information content of the motif	http://meme.sdsc.edu/	7584439
MITRA	Uses an efficient data strauctture to traverse the space of IUPAC patterns	http://www.calit2.net/compbio/mitra	12169566
MotifSampler	Matrix-based , motif-finding algorithm that extends Gibbs sampling by modeling the background with a higher order Markov model	http://www.esat.kuleuven.acbe/~dna/Biol/Software.html	11751219
Oligo/dyad-analysis	Detects overrepresented oligonucleotides with oligo-analysis and spaced motifs with dyad-analysis	http://rsat.scmbb.ulb.ac.be/rsat	10734201
SeSiMCMC	Modification of Gibbs sampler algorithm that models the motif as a weight matrix, optionally with the symmmetry of a palindrome or of a direct repeat, and optionally with spacers	http://favorov.hole.ru/gibbslfm/	15728117
Weeder	Consensus-based method that enumerates exhaustively all the oligos up to a maximum length and collects their ocuurences (with substitutions) from input sequences	http://159.149.109.16/Toll/ind.php	15215380
YMF	Uses an exhaustive search algorithm to find motifs with the greatest z-scores	http://bio.cs.washington.edu/software.html#ymf	12824371

Dragon Motif Builder (DMB) (E Huang *et al.*, 2005) is based on the Expectation Maximization (EM) algorithm. The Expectation maximization (EM) algorithm estimates the maximum likelihood of parameters in probabilistic models, where the model depends on unobserved (latent) variables. EM alternates between performing an expectation (E) step, which computes the expected value of the latent variables, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters given the data and setting the latent variables to their expectation.

In DMB, EM is used to estimate the probability density of the most popular patterns within a set of DNA sequences. The optimal motifs are predicted with pattern matching score function and the population of the motifs among the sequences. The EM algorithm iteratively augments the motif data by guessing the values of the optimal score and population with the sequence, and then re-estimates the parameters by assuming the "best" value for the motif group. In order to model the probability density of the data effectively, most likelihood function was implemented to choose the initial value that has highest converged likelihood value. The threshold coefficient for information content has been applied to improve the efficiency and accuracy of the search approach.

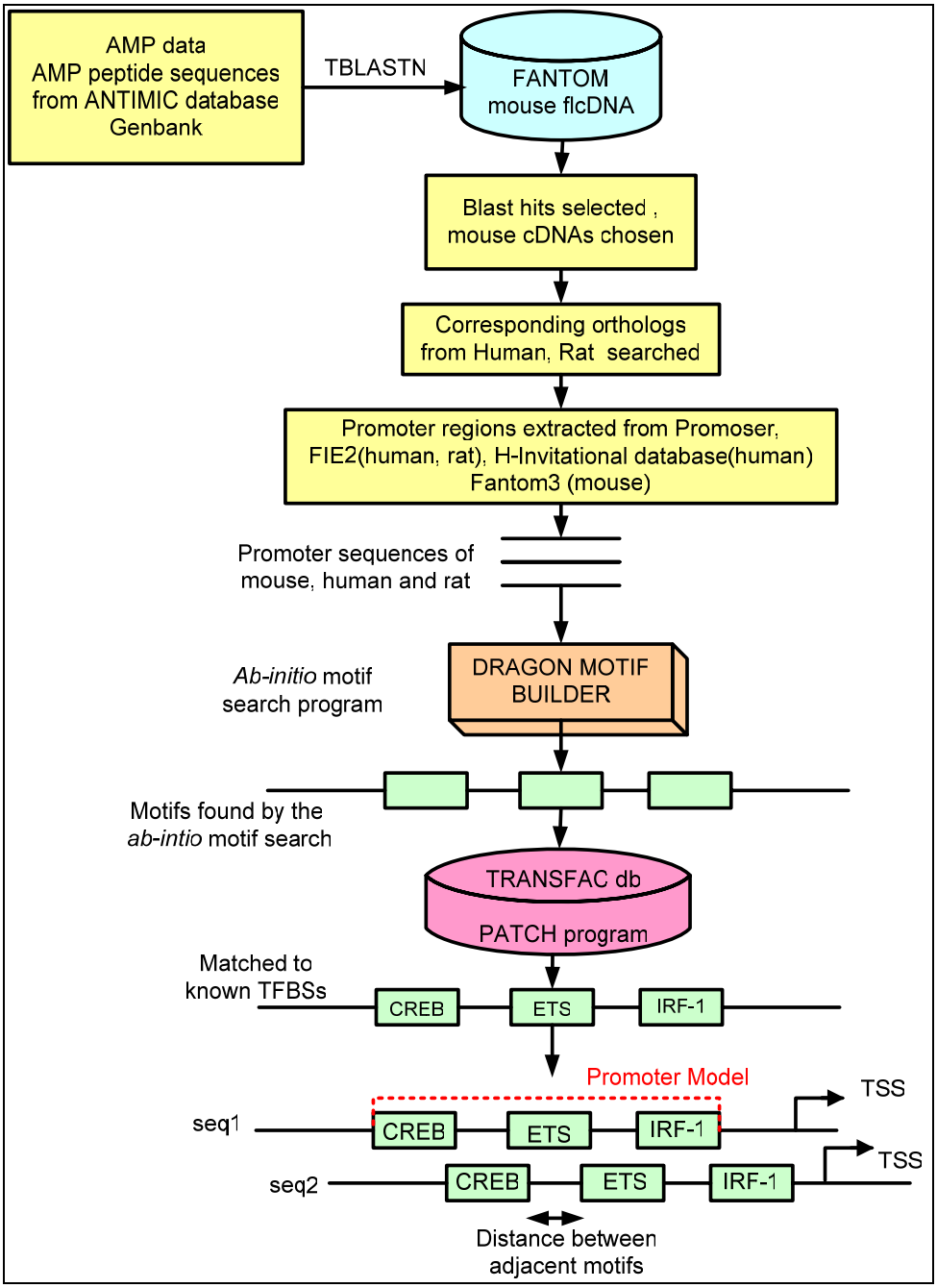
Figure 5.2: Schematic representation of the DMB algorithm



5.3 Materials and methods

The strategy used to find the motifs in AMP promoter regions is schematically depicted in **Figure 5.2**. Most of the AMP sequences were extracted from the ANTIMIC database (Brahmachary *et al.*, 2004) (http://www.research.i2r.org.sg/Templar/DB/ANTIMIC/) that contains the largest number of non-redundant AMPs (1,439) and GenBank. TBLASTN (Altschul *et al.*, 1990) with BLOSUM45 matrix was used to search 102,801 flcDNAs of the FANTOM collection (Carninci *et al.*, 2005) (FANTOM1+2 (60,770) plus FANTOM3 (42,031)) against AMP protein sequences of ANTIMIC. Since TBLASTN translates the query sequence into six possible open-reading frames, cDNAs with short CDS below the protein-coding annotation threshold can be captured. From the translated flcDNA sequence out of 183 mouse candidates with sequence identities to known AMPs equal or greater than 60% over length of 100 residues or with E-value of 0.01 or less, five were identified as false positives by checking their stable gene name and gene ontology annotations. Less stringent threshold settings (i.e. 50% or 55%) applied to a test set of cathelicidins, alpha and beta defenins led to too many false positives (data not shown) without gaining any new AMPcg candidates among the FANTOM sequence set.

Figure 5.3: Workflow of promoter sequence set preparation and analysis



AMP peptide sequences were collected from ANTIMIC and Genbank databases and searched with TBLASTN against FANTOM3 cDNA sequences applying a cut-off of equal or greater than 60% identity. The promoter regions [-1000, +200 nt] of mouse AMPcg, human and rat orthologs were extracted and submitted to Dragon Motif Builder (DMB) for *ab initio* motif searching. The resulting consensus motifs were passed to TRANSFAC and compared with known TFBSs using the PATCH program

5.3.2 Extraction of promoter regions

The mouse flcDNA were annotated with their official gene names and symbols, associated representative cDNAs, chromosomal localization information, TUID (transcriptional unit ID) and CAGE TSS (transcription start site information based on CAGE tags) (Carninci *et al.*, 2005). Human and rat orthologs were determined for the AMP-coding mouse flcDNAs, using the Entrez Gene (Maglott *et al.*, 2005) and HomoloGene (Wheeler *et al.*, 2005). In addition, each of these ortholog groups was manually checked for synteny. The promoter regions of the orthologs in human and rat were extracted using PromoSer (http://biowulf.bu.edu/zlab/PromoSer) (Halees *et al.*, 2003) and FIE2 (http://research.i2r.a-star.edu.sg/promoter/FIE2_1) (Chong *et al.*, 2003, Halees *et al.*, 2003) programs, as well as H-Invitational database (Fujii *et al.*, 2004). All three resources provide estimated TSS locations based on mapping EST and flcDNA data to genomic sequences. The promoter regions extracted for mouse, human and rat covered (-1000, +200) relative to the estimated transcription start site (TSS) location. In the case of multiple TSS locations in human and rat sequences the most 5' one was extracted. The TSS location of mouse sequences was determined by using the start position of the first exon of the FANTOM cDNA-genome mapping data (http://fantom31p.gsc.riken.jp/cage/download/mm5/cage.rep_tag.2004-11-16.chr_all_gff.tar.gz). Mouse promoter sequences (-1000, +200) were then extracted by mapping the TSS location to the mouse genome data from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/mm5/chromosomes/). The final dataset contained 77 promoters from mouse, rat and human. Only seven mouse sequences had associated CAGE tag information (**Supplementary Table 5.1**). Therefore, TSS location

was estimated for all sequences based on the 5'end of the flcDNA data. For histone2a genes a region of (-200, +100) relative to the TSS was extracted because these genes appear to have bidirectional promoters within 200 nt of the TSS.

5.3.2 Motif search

The promoter sequences were submitted to the Dragon Motif Builder (DMB) program (E Huang *et al.*, 2005) (http://research.i2r.a-star.edu.sg/DRAGON/Motif_Search/) for *ab-initio* motif finding. The EM threshold was set to 0.85 for all families that lacked experimentally confirmed TFBSs in their promoters. One should note that there is no rule about what is the optimal threshold. In fact, the optimal threshold is likely to be different for different promoter sets. Thus, a somewhat arbitrary threshold of 0.85 was used because it resulted in relatively specific matrix families. Since the algorithm is heuristic, different thresholds usually produce different results. In the cases when there have been known functional TFBSs, for the AMpcg family, two different thresholds (0.85, 0.75) were used and the one selected was the one that fitted better to the experimentally confirmed TFBSs, as this would very roughly approximate selection of a more optimal threshold in these cases. The program was set to search for 20 motif families, with motifs of length 10 to 15 nt within each of the 22 AMPcg families. In total 440 motif families were identified. In the case of the histone2a family a shorter motif length of 8-12 nt was chosen because the promoter length of histone2a family was shorter than for the other families. After DMB identified the sequence motifs, Patch program (mismatch =0; motif length =6; species =all) was used (Wingender *et al.*, 2000) of TRANSFAC professional database ver. 8.4 to infer potential transcription factors (TFs) that may bind to motifs of

these families. Promoter models were created from motifs that were conserved among the all promoter sequences of an AMPcg family.

To find motif families that are common across many AMPcg families, all 440 motifs were combined and searched for the most commonly found sub-motif families in them. For this the DMB program was used for searching for motifs of 6-8 bp length. The reduction of motif length did not cause over-prediction of motifs since the search was restricted to sequences of the previously identified motifs of length 10-15 bp. Potential motif-binding TFs were identified by the Patch program as already described.

For the penk family, three programs DMB, MEME (http://meme.sdsc.edu) (Bailey and Elkan, 1994) and Improbiser (http://www.soe.ucsc.edu/~kent/improbizer) were used to search for motifs of 10-15 nt length, based on EM algorithm. Improbiser can identify a maximum of six motif families. For MEME and DMB 20 motif families were identified and the top six families based on e-value, were selected. This threshold setting allowed us to obtain comparable results from three different programs. The motifs were then compared with TRANSFAC database entries to obtain TFs that can potentially bind to these motifs. **Figure 5.3** shows the workflow of *ab-inito* based motif finding.

5.3.3 Phylogenetic analysis

Multiple sequence alignments and phylogentic analyses of alpha-defensin sequence were done using Clustal W (Thompson *et al.*, 1994) and MEGA3.0 (Kumar *et al.*, 2004). Alpha-defensin sequences covering (-1000, +200) region relative to TSS were extracted using the Ensembl (Birney *et al.*, 2004) gene data export function. UPGMA (unweighted pair group method with arithmetic averages) phylogenetic trees for alpha-defensins were

constructed with Mega3.0 (Kumar *et al.*, 2004) using Kimura 2-parameter methods with 1000 bootstrap replications.

5.3.4 Statistical significance of potential NHR-binding motifs

All families were sorted according to the number of motifs that may bind NHR. Then, the author split AMPcg families into two groups, A and B. In group B the family that had the least number of such motifs was included. The remaining families were placed in group A. P-value was calculated for the enrichment in motifs that may bind NHR. The p-value is determined using the hypergeometric distribution and the right-side Fisher's exact test and was corrected by the Bonferroni method for the 440 tests (this is the number of motif families identified; 20 motif families for each of the 22 AMPcg families). The author then excluded from group A the AMPcg family with the next least number of target motifs and added that family to group B. The P-value calculation was repeated. This process of eliminating AMPcg families from group A is repeated until A contained the last of the 22 AMPcg families. Based on the 21 p-values calculated this way (**Supplementary Table 5.7**), the one with the smallest value was determined, 2.81167E-06 (Bonferroni corrected value = 0.001237134). This determines the group of 11 AMPcg families that are significantly enriched by motifs that potentially bind NHRs.

5.4 Results and discussion

5.4.1 Novel AMP transcripts

The FANTOM3 data set comprising more than 100,000 flcDNAs has recently been released (Carninci *et al.*, 2005). Macrophages, cells of the innate immune system, were a major source of additional new flcDNAs in this set. ANTIMIC-derived AMP sequences were mapped to the FANTOM3 cDNA set to search by sequence similarity (TBLASTN) for candidate cDNAs encoding new members of AMP families. Of 183 mouse candidates with sequence identities to known AMPs equal or greater than 60% over length of 100 residues or with E-value of 0.01 or less, five were identified as false positives by checking their stable gene name and gene ontology annotations. Thus, 178 AMPcg sequences belonging to 29 families were identified. One hundred and three new mouse transcripts belonging to the AMP families alpha-defensin, alpha2casein, apoa2, beta-defensin, spag11, bpi, calgranulin, cathelicidin, cathepsinG, dbi, slpi, enhancer of rudimentary homolog, granulin, hepcidin, histone2a, IFN-inducible antiviral protein Mx, lactoferrin, lysozyme, mbp, melanotropin alpha, ovotransferrin, proenkeph alin 1, sap2, secretogranin, skiv2l, spyy, vasostatin, vip and zap, were found in the FANTOM3 (without FANTOM1+2) sequence subset. All new members were sequenced from cDNA libraries of immune cells (i.e. macrophages), adipocytes and testis, among others, indicating that the transcriptome of inducible genes involved in innate immunity is still incomplete.

The definition of true orthologies across species is difficult in multigene families associated with innate immunity, wherein gene duplication is a common feature of evolution even within the mammalia. For example, a review of the S100 (calgranulin)

family noted that there are three members of the myeloid-associated family (S100A8, A9 and A12) in humans, but only two (S100A8 and A9) in mice (Ravasi *et al.*, 2004). Correspondingly, it was found that the mouse AMP casein delta (csnd), defensin-related sequence cryptidin peptide (Defcr-rs1), mast cell protease family (mcpt2, mcpt4, mcpt8), and histone2a (Hist2h2aa2), did not have corresponding family members in human (**Supplementary Table 5.2**). On the other hand, the Rnase A family member Rnase 7 was found in human, but was absent in mouse. Within the beta-defensin and alpha-defensin family members, cDNA sequences confirm mouse-specific expansion reported in previous genome-based studies (Schutte *et al.*, 2002 and Scheetz *et al.*, 2002).

The analysis was restricted to the three mammalian species as the approach was aimed at finding differences and similarities in mammalian orthologs of mouse data from the FANTOM3 project. Orthologs of mouse genes in invertebrates and cold-blooded vertebrates are too distant for such promoter analysis. Another problem is the absence of very accurate promoter data sets for these species, which are necessary for this type of analyses. This resulted in consideration of only a subset of *bona fide* orthologous mouse, human and rat promoter sequences representing only 22 out of 29 AMP families. . For these 22 AMP families, 31 promoter regions from mouse with the corresponding 30 and 15 promoter orthologs from human and rat, respectively, were extracted (Supplementary **Table 5.1**). Mouse cryptidins were included in the alpha-defensin family because they represent a subfamily of alpha-defensins (Eckmann, 2005). The analyzed families and the sequence accessions of their members are listed in **Supplementary Table 5.1**

5.4.2 Promoters and ab-initio motif discovery

Having assembled a set of candidate AMPcgs, the aim was to identify sets of potential transcriptional control elements common to all or some of these genes. For many of these genes, the precise TSSs have been identified through the high throughput CAGE technology, since macrophages were extensively polled with this method (Carninci *et al.*, 2005). However, for this thesis, (-1000 to +200) promoter region relative to the longest cDNA was chosen. The most common current approach to identification of motif complements amongst co-regulated genes is to search using predetermined position-weight matrices for known TFBSs as available from TRANSFAC, JASPAR and other sources. This approach presumes that binding is not influenced by context. The author has used the ab-inito approach for finding TFBS motifs. . There are several *ab-initio* motif discovery programs available (Tompa *et al.*, 2005). No program shows a distinct advantage over others on all data types. However, the author compared the performances of DMB (E Huang *et al.*, 2005), an in-house developed program with two other programs, MEME (Bailey and Elkan, 1995) and Improbiser (http://www.soe.ucsc.edu/~kent/improbizer/improbizer.html). All three programs use *ab-inito* motif discovery algorithms based on Expectation Maximization. Promoter sequences of the proenkephalin (penk) AMP group (4922504O09, HIX0007519.2, NM_017139) were used, which has been studied empirically in transfection assays. Penk promoters are known to possess a TATA box and respond to cyclic AMP, glucocorticoids and protein kinase C (AP1) agonists (Kobierski *et al.*, 1999 , Garcia-Garcia *et al.*, 1998 , Fu *et al.*, 1997). Since Improbiser can identify only six motifs, the top six motifs produced by each of these systems were considered first. Among the top six motifs,

DMB- reported three motifs (TATA, AP-2, AP-1) that may bind TFs known to control the penk promoter (Fu *et al.*, 1997 , Le *et al.*, 2003). MEME reported one motif (TATA) and Improbiser two (NF-Y, TATA) motifs. As DMB and MEME can identify arbitrary number of motifs, top 20 motifs generated by DMB and MEME were considered. Seven DMB-derived motifs coincided with known TFBSs (TATA, NF-kappaB, AP-2, AP-1 NFI/CTF, NF-Y, MZF1, MIG1, MBP-1) (Fu *et al.*, 1997, Le *et al.*, 2003) known to control the penk promoter. MEME yielded only three known penk promoter motifs (TATA, NFI/CTF, AP-1). Considering the differences in performance and the longer computation time of MEME, DMB was the preferred program for the entire analysis.

5.4.3 Phylogenetic analysis of defensins

Alpha-defensins are specific to mammals. Phylogenetic analyses of alpha-defensin protein-coding sequences was done to provide support for gene duplication events and rapid evolution under positive selection pressure (Patil *et al.*, 2004). Gene duplication events have probably led to both species-specific and functionally diverse subsets of alpha-defensins, which should be also detectable in the upstream regulatory regions. The author was interested to see how promoter content reflects phylogenetic similarity. Nine alpha-defensin promoters of mouse, rat, chimpanzee and human were analyzed in terms of phylogenetic, functional and motif relationships. The UPGMA tree (**Figure 5.4**) shows two clusters. With the exception of rat Defcr4, the tree topology coincides with the previously reported (Patil *et al.*, 2004) enteric (i.e. intestine) and myeloid/neutrophil cell expression of rat, mouse and human alpha-defensins. Enteric-expressed defensins are important to barrier function of the gut mucosal surface against bacteria, whereas

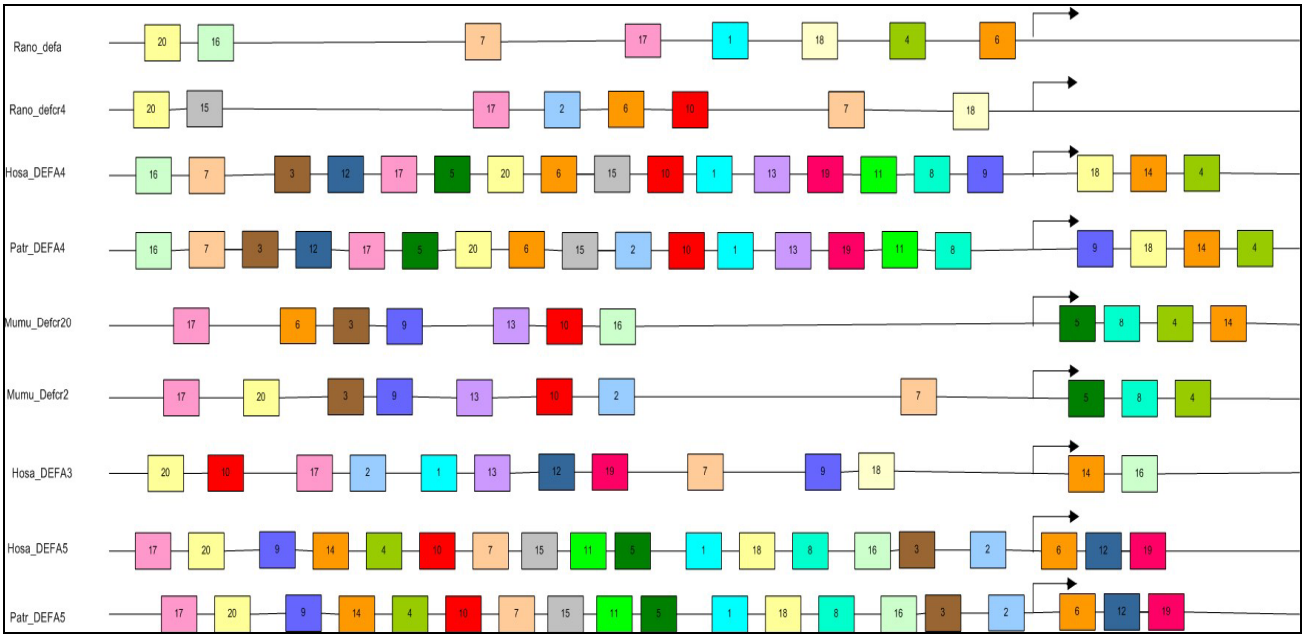
myeloid and neutrophil-specific defensins help macrophages and neutrophils to kill internalized bacteria. Human DEFA3, DEFA4, chimpanzee DEFA4 and rat Defa represent the myeloid-specific alpha-defensins. Mouse Defcr20, Defcr2, rat Defcr4, human and chimpanzee defa5 belong to the enteric-expressed group of alpha-defensins. Comparison of myeloid-expressed rat Defa with enteric-expressed mouse defcr2 promoter regions showed that the common arrangement 20-7-4 of three promoter motifs was conserved in rodents (**Figure 5.4**). The annotation means that in the promoter region from 5' to 3' the order of motifs identified is: 'motif20 – motif7 – motif4'. These three motifs potentially may bind: motif20 (AR PXR-1: RXR-alpha), motif7 (POU1F1a, POU2F1), motif4 (RAR-alpha1, RXR-alpha) (**Table 5.3)**. This arrangement of motifs appears to be unique to mouse Defcr2 and rat Defa and thus, suggests association with the specific myeloid or enteric expression.

A comparison of myeloid–specific human and chimpanzee sequences (Hosa_DEFA4, Patr_DEFA4, Hosa_DEFA3) and enteric sequences (Hosa_DEFA5 and Patr_DEFA5) showed that they share arrangement of four motifs (20-10-11-19). For myeloid-specific primate sequences Hosa_DEFA4, Patr_DEFA4, Hosa_DEFA3 a common arrangement of eight motifs (20-10-1-13-1-19-9-18-14) was found across the promoters. When myeloid-specific rat sequence Rano_DEF1 was included, only a common arrangement of three motifs (17-1-18) was found, that was specific to all myeloid sequences in the data set. Enteric primate sequences (Mumu_Defcr20, Mumu_Defcr2, Rano_Defcr4, Hosa_DEFA5, Patr_DEFA5) have an arrangement of three motifs (17-10-7) in common. Motif 17 (GMASTTCTKT) was found common between all the myeloid and enteric sequences in the data set. This motif contains a sub-motif that

is a putative binding site for IRF-1, IRF-3, NF-AT1, NF-AT2, NF-AT3, NF-AT4. In the case of rodent sequences (Rano_DEF1, Rano_DEFCR4) motifs 20, 7 and 16, associated with putative binding sites for YY1, STAT5A, IRF-1, IRF-3, NF-AT1, NF-AT2, NF-AT3, NF-AT4 were common between the two sequences (**Figure 5.4**, **Table 5.3**, **Supplementary Figure 5.1**). The upstream regulatory regions of mouse cryptidin alpha-defensins contained eight common motifs with similar positioning.

The extremely low sequence homology of beta-defensin promoters of orthologs and paralogs among cow, mouse, rat, chimpanzee and human, together with different exon-intron structures suggests multiple events of functional changes or acquisition of new functions as a result of positive diversifying selection during evolution (Maxwell *et al.*, 2003, Morrison *et al.*, 2003, Semple *et al.*, 2003. Additional analysis with RepeatMasker (http://www.repeatmasker.org) also revealed various retro-transposons in the upstream regions of rat and mouse beta-defensins that are absent in primates. Probably the most striking example of functional specialization in the primate lineage is SPAG11. SPAG11 is derived from the ancestral fusion of two beta-defensins. Expression of SPAG11 AMPs appear to be androgen-dependent and restricted to the male urogenital tract (Avellar *et al.*, 2004).

Figure. 5.4 Motif distribution in alpha-defensin promoters



The boxes represent the motifs found by *ab-initio* searching. The numbers (i.e. 13) in the boxes refer to different motifs. The grey line connecting the boxes denotes a promoter region of 1,200 bp length. The broken arrow indicates the TSS. The species abbreviations are Rano: *Rattus norvegicus*, Mumu: *Mus musculus*; Patr: *Pan troglodytes*; Hosa: *Homo sapiens*.

Table 5.3 Common motifs found between groups of enteric and myeloid-specific alpha-defensin sequences

The species abbreviations are Rano: *Rattus norvegicus*, Mumu: *Mus musculus*; Patr: *Pan troglodytes*; Hosa: *Homo sapiens*. Unknown: motif does not match any of the TRANSFAC-listed TF binding sites.

Common Motif	Consensus Motif	Putative TFBS	Gene name
20	AGAARCTCAGS	AR, PXR-1:RXR-alpha	*Hosa_defa4* (myeloid), *Patr_defa4* (myeloid), *Hosa_defa3* (myeloid), *Patr_defa5* (enteric)
10	CATAMTACCTGA	AP-1, c-Jun	*Hosa_defa4* (myeloid), *Patr_defa4* (myeloid), *Hosa_defa3* (myeloid), *Patr_defa5*(enteric)
11	KAGYTTTTWTCC	GATA-1, NF-AT1,NF-AT2,GATA-6,GATA-3,NF-AT3,NF-AT4	*Hosa_defa4*(myeloid), *Patr_defa4* (myeloid), *Hosa_defa3*(myeloid), *Patr_defa5*(enteric)
19	AGTAAAGCCA	Unknown	*Hosa_defa4*(myeloid), *Patr_defa4* (myeloid), *Hosa_defa3*(myeloid), *Patr_defa5*(enteric)
20	AGAARCTCAGS	YY1 STAT5A	*Rano_DEF1* (myeloid), *Rano_DEFCR4* (enteric)
17	GMASTTCTKT	IRF-1 IRF-3 NF-AT1 NF-AT2 NF-AT3 NF-AT4	*Rano_DEF1* (myeloid), *Rano_DEFCR4* (enteric)
6	GAAAAAAGAAT	Unknown	*Rano_DEF1* (myeloid), *Rano_DEFCR4* (enteric)
20	AGAARCTCAGS	AR PXR-1:RXR-alpha	*Rano_DEF1* (myeloid), *Mumu_Defcr2* (enteric)
7	AAAMATYCAT	POU1F1a, POU2F1	*Rano_DEF1* (myeloid), *Mumu_Defcr2* (enteric)
4	GAAGGACCAGC	RAR-alpha1, RXR-alpha	*Rano_DEF1* (myeloid), *Mumu_Defcr2* (enteric)
17	GMASTTCTKT	GR AR	*Mumu_Defcr20* (enteric), *Mumu_Defcr2* (enteric)
3	ATTCTCHTGGACA	RXR-beta T3R-alpha1 T3R-beta1 USF1b USF1 GR	*Mumu_Defcr20* (enteric), *Mumu_Defcr2* (enteric)
9	CTCTTGCCTG	C/EBPalpha	*Mumu_Defcr20* (enteric), *Mumu_Defcr2* (enteric)
13	GGAATCAAGT	Unknown	*Mumu_Defcr20* (enteric), *Mumu_Defcr2* (enteric)
10	CATAMTACCTGA	AP-1 c-Jun	*Mumu_Defcr20* (enteric), *Mumu_Defcr2* (enteric)
5	CCTGCTCCCTGBT	AR T3R-alpha RXR-alpha VDR	*Mumu_Defcr20* (enteric), *Mumu_Defcr2* (enteric)
8	TGTCCTGGTCC	GR PR-alpha PR-beta PR B RAR-alpha1 RXR-beta RAR-gamma T3R-alpha T3R-beta1 T3R-beta2 HNF-4alpha RAR-alpha RAR-alpha:RXR-gamma ,RAR-beta RAR-beta:RXR-alpha AR NFI/CTF RXR-alpha VDR ERR1	*Mumu_Defcr20* (enteric), *Mumu_Defcr2* (enteric)
			Mumu_Defcr20 (enteric), *Mumu_Defcr2* (enteric)
			Mumu_Defcr20 (enteric), *Mumu_Defcr2* (enteric)
4	GAAGGACCAGC	RAR-alpha1 RXR-alpha	*Mumu_Defcr20* (enteric), *Mumu_Defcr2* (enteric)

5.4.4 Frequently occurring TF binding motifs conserved across many AMPcg families

The transcriptional regulation of AMPcg families varies from family to family because of the different AMP characteristics and their tissue cell-specific expression. Thus, it would not be expected that different AMPcg families share considerable similarities in their promoters. It is thus challenging to explore if such similarities exist and whether a TF can be involved in the control of more than one AMPcg family. In this section, the author looked for possible common motifs that are shared across several AMPcg families (see Methods). By analyzing only those motif groups that were found to be shared by at least six AMPcg families, eight motif groups out of total of 94 motif instances were found from 31 mouse, 30 human and 15 rat AMP promoter sequences (**Table 5.4**). This suggests that, a core set of TFs exist that participate in transcription activation of many examined AMPcg families in all three examined species.

Species-specific differences were observed in the combination of motifs and positions relative to the TSS. Each of the motif families is represented by the consensus motif obtained from all motif instances in that family. The consensus motif AGGAAA is known to be recognized by TFs PEA3, c-Ets1, E74A, PU.1, LyF-1, c-Ets-2, ISGF-3, NF-AT1, NF-AT2, NF-AT4 and DEAF-1. Consensus motifs ACAGCA and ATGGAG are specific for GR and Nkx2-1, respectively. Consensus motif CCCGCCCC corresponds to binding site for TFs Sp1/Sp3. TGGCATT recognizes TF NF-1. CCAGGG, ACCTGG and TCTTTC did not match to any known TFBS contained in the TRANSFAC database. These three consensus motifs could represent potentially novel *cis*-elements.

Comparison of the TFs associated with the predicted motifs showed that four consensus motifs correspond to the published experimentally confirmed TFs of AMPcgs,

such as, GR for motif ACAGCA. This motif was conserved among 32 genes of ten different AMP families in mouse, rat and human. PEA3, c-Ets1, PU.1, LyF-1, c-Ets-2, NF-AT1, NF-AT2 and NF-AT4-specific motif AGGAAA was observed in 34 genes belonging to 11 AMPcg families. Sp1 and Sp3-specifc motif CCCGCCCC appeared in 15 genes derived from six AMPcg families. NF-1 motif TGGCATT was present in 36 genes of nine AMP families. All these motifs were found in human, mouse and rat AMP genes (see **Table 5.4**). Consensus motif CCAGGG was observed in 24 genes of eight AMPcg families. ACCTGG was present in 28 genes of seven AMPcg families. TCTTTC motif occurred in 26 genes of nine AMPcg families.

Four motifs appeared to be species- or lineage-specific. For example, AGGAAA motif was found only in three rodent genes of the lysozyme family. CCAGGG was absent in genes of the human Spag11 family. TGGCATT motif was absent in human genes of the Apoa2 and Spyy families. CCCGCCCC was not found in mouse genes of the Apoa2 family (**Table 5.4**). However, one should note that this observation has been made for the region of (-1000, +200) bp of the promoters. Similar species-specific differences were reported for the promoter of mouse and human Toll-like receptor 3 and its expression pattern (Heinz *et al.*, 2003). It is possible that these AMP genes are regulated by different promoter regions in mouse and human. Due to lack of sufficient data on microbial context, signaling pathways and TF binding-data on AMPs, it remains to be seen whether these disparities reflect an exposure to a different microbe environment or a physiological differences. Thus, it can be concluded that in spite of differences in functions of AMPcg families and differences in their tissue cell-specific expression, their promoters share a number of common motifs. It is also possible that at least some of

these common motifs function as binding sites for unknown or undiscovered TFs. These motifs also represent interesting experimental targets, at least for the assessment of binding of TFs suggested through the computational analysis.

Among the most frequently occurring motifs in promoters of AMPcgs the analysis identified three PEs (CCAGGG, ACCTGG and TCTTTC) that are not known to bind any of the TRANSFAC contained TFs and thus could likely represent novel *cis*-elements.

Table 5.4: Motifs that are highly enriched among different AMP families.

Pattern: The consensus of motif sequence found in the AMP sequences; TF name: Transcription Factor name associated with the motif; Total AMP: The number of AMP families that contain the motif; AMP family: The AMP families which contain the motif; Seq IDs: The mRNA ids of AMPcgs whose promoter sequences are analyzed; Mm: mouse, Hs: human, Rn: rat (if the motif is found in a species it is denoted by "+" else it is denoted by "-").

No.	Pattern	TF name	Total AMP	AMP Family	Seq IDs	Mm	Hs	Rn
1	ACAGCA	GR	10	Alpha defensin	2010016B13, 2010016F14, NM_021010, NM_001926, NM_001925, NM_005217	+	+	-
				Apoa2	I530003A11,HIT000032344.2, NM_013112	+	+	+
				BPI	9230105K17,BC040955	+	+	-
				Calgranulin	F430201H11, NM_002965, NM_053587	+	+	+
				Hepcidin	NM_052971,2210420P15	+	+	-
				Histone 2A	9030420B16, NM_003512, 1190022L06, NM_021052	+	+	-
				Melanotropin alpha	5730403F20, NM_000939, NM_139326	+	+	+
				Secretogranin	5730420J08,HIX0015625.2, NM_012526	+	+	+
				Vasostatin	G630083O06,HIX0011909.2,NM_021655	+	+	+
				ZAP	F420004O17,HIX0007129.3,NM_173045	+	+	+
2	AGGAAA	PEA3, c-Ets1, E74A, PU.1, LyF-1, c-Ets-2, ISGF-3, NF-AT1, NF-AT2, NF-AT4,	11	Alpha defensin	2010016B13, 2010016F14, NM_021010, NM_001926, NM_001925, NM_005217	+	+	-

		DEAF-1						
				BPI	9230105K17,BC040955	+	+	-
				Calgranulin	F430201H11, NM_002965, NM_053587	+	+	+
				Cathelicidin	F930015N03,NM_004345,AF484553	+	+	+
				Hepcidin	2210420P15,NM_052971	+	+	-
				Histone 2A	9030420B16,NM_003512,11900 22L06,NM_021052	+	+	-
				Lysozyme	9530003J23,I420013M05, NM_012771	+	-	+
				MBP	2510004C07,HIX0009634.2,NM _031619	+	+	+
				Proenkaphalin	4922504O09,HIX0007519.2,NM _017139	+	+	+
				Secretogranin	5730420J08,HIX0015625.2, NM_012526	+	+	+
				VIP	9130007F05,HIX0006306.2	+	+	-
3	CCAGGG	unknown	8	Alpha defensin	2010016B13,2010016F14, NM_021010,NM_001925	+	+	-
				Spag11	9230111C08,NM_145087	+	-	+
				BPI	9230105K17,BC040955	+	+	-
				DBI	6720460E16,NM_020548	+	+	-
				Granulin	0610012H06,BC000324, NM_017113	+	+	+
				Lysozyme	9530003J23,I420013M05, AF099029,NM_012771	+	+	+
				Melanotropin alpha	5730403F20,NM_000939,NM_1 39326	+	+	+
				SPYY	0710005A05,C820007C10, HIX0006525.2,NM_012614	+	+	+

4	ACCTGG	unknown	7	Betadefensin	9230107O10, AF525930, D630029A12, BC033298, NM_031810, NM_153324, 2310001F05, NM_004942, 1700011J22, NM_152250, 9230103N16, 4930563B01	+	+	+
				BPI	9230105K17,BC040955	+	+	-
				Calgranulin	F430201H11, NM_002965, NM_053587	+	+	+
				Cathelicidin	F930015N03,NM_004345,AF484553	+	+	+
				Granulin	0610012H06,BC000324, NM_017113	+	+	+
				Lactoferrin	9830118D19,NM_002343	+	+	-
				ZAP	F420004O17,HIX0007129.3,NM_173045	+	+	+
5	ATGGAG	Nkx2-1	10	Alpha defensin	2010016B13,2010016F14, NM_001926,NM_001925,NM_005217	+	+	-
				Calgranulin	F430201H11, NM_002965, NM_053587	+	+	+
				Cathelicidin	F930015N03,NM_004345,AF484553	+	+	+
				DBI	6720460E16,NM_020548	+	+	-
				Slpi	2310075E18,HIT000038907.2,NM_053372	+	+	+
				Hepcidin	2210420P15,NM_052971	+	+	-
				Lactoferrin	9830118D19,NM_002343	+	+	-
				MBP	2510004C07,HIX0009634.2,NM_031619	+	+	+
				VIP	9130007F05,HIX0006306.2	+	+	-
				Vasostatin	G630083O06,HIX0011909.2,NM_021655	+	+	+
6	TCTTTC	unknown	9	Alpha defensin	2010016B13,2010016F14, NM_001925,NM_005217	+	+	-
				BPI	9230105K17,BC040955	+	+	-
				Calgranulin	F430201H11, NM_002965, NM_053587	+	+	+

				Slpi	2310075E18,HIT000038907.2,NM_053372	+	+	+
				Hepcidin	2210420P15,NM_052971	+	+	-
				MBP	2510004C07,HIX0009634.2,NM_031619	+	+	+
				Melanotropin alpha	5730403F20,NM_000939,NM_139326	+	+	+
				SPYY	0710005A05,HIX0006525.2,NM_012614	+	+	+
				ZAP	F420004O17,HIX0007129.3,NM_173045	+	+	+
7	CCCGCCCC	Sp1, Sp3	6	Alpha defensin	2010016B13,2010016F14, NM_021010,NM_001926	+	+	-
				Apoa2	HIT000032344.2,NM_013112	-	+	+
				BPI	9230105K17,BC040955	+	+	-

5.4.5 Distribution of known TF binding motifs in AMPcg families

Prior to discussing results of specific AMPcg families the predicted motifs were compared with the experimentally verified motifs documented in previous reports. The predicted motifs comprise binding sites for various immune-response related TFs (e.g. NF-kappaB) and nuclear hormone receptors (i.e. RXR alpha). The *ab-initio* determined motifs potentially bind 41 (59%) out of 70 experimentally confirmed TFs that participate in the control of these AMPcg families. Among members of the lactoferrin family, all experimentally reported TFs (SP1, C/EBP) were found. Six AMP families (zap, apoa2, calgranulin, granulin, spyy, bin1b/spag11) lacked published experimental information on associated TFs. DMB-predicted motifs for these families include 57 motifs conserved among mouse and human. **Supplementary Table 5.3** shows a comparison between the motifs that were found by *ab-initio* approach versus those reported to be experimentally found for each of the AMPcg families. The list of experimentally detected TFs that is presented here is not exhaustive, but it well supports the *ab-initio* motif finding method. For each AMP family, motifs were found that did not match any of the known TRANSFAC-contained motifs and were reported as "unknown motifs". Other set of motifs matched to known TFBS but were previously not reported to control AMPcgs. These new AMPcg-associated candidates are shown in **Supplementary Table 5.4**.

In this section, all the predicted TFs that potentially bind motifs identified for different AMPcg families have been categorized into ten tissue specific categories and two general categories of cell-cycle specific TFs and nuclear hormone receptors (NHRs). This work was done in collaboration with VB Bajic. **Table 5.5** and **Supplementary Table 5.5** show distribution of motifs identified by DMB across all AMP families. The

motifs were compared to TRANSFAC-contained motifs to determine their correspondence with the known TFBSs. Twelve different categories of TFs were considered. These are adipocyte-related, NHR, cell cycle-related, immune cell-specific, liver cell-specific, lung cell-specific, muscle cell-specific, nervous system-related, pancreatic beta cell-specific, pituitary gland-specific, eye-specific, and bone- (and teeth-) specific TFs. The categories were chosen based on supporting knowledge of links, for example between the immune system and a particular category. For example, microglia cells which are brain macrophages (Moran *et al.*, 2004) would represent a link between the nervous system and immune system. The association of TFs with different TF groups is based on the TRANSFAC database collections and literature survey.

To determine the dominant TF categories that are potentially involved in control of AMPcg families, motifs that TFs could bind to were analyzed, as well as the distribution of the TFs across the 22 AMPcg families. For each of the AMPcg family only the top two-ranked TF categories were considered. The ranking was based on the proportion of motifs that potentially bind TFs of specific category in any AMPcg family. Cases were considered when TF-binding motifs associated with a particular TF category occurred in 25%, 30%, 35% or 40% of all motifs observed in an AMPcg family. Three TF categories (liver-specific, neuron system-specific, NHR) appeared to be either the first or second ranked in three out of four considered cases, and these TF categories, also represent the top ranked ones, overall. The results are summarized in **Table 5.5**. The appearance of NHRs in these top ranked TF groups is unexpected. If we require that at least 35% (7 out of 20) of the identified motifs for each of the AMPcg families can bind TFs from a particular group, NHR and neuron system specific TFs appear in 11 out of 22

AMPcg families. The next one is the group of liver-specific TFs (10 families), followed by adipocyte-specific TFs (8 families) and immune cell-specific TFs (5 families).

Further, the group of AMPcg families were determined that are most enriched in motifs that potentially bind NHRs. This group is determined as explained in Methods and contains 11 AMPcg families. Each of the families contains at least 35% (7 out of 20) of motifs that potentially bind NHRs. These 11 families are alpha-defensin, lactoferrin, hepcidin, bin1b, zap, dbi, cathelicidin, proenkaphalin, mbp, slpi, bpi. The statistical significance of the enrichment of NHR related motifs in this group is based on the Bonferroni corrected p-value obtained from the right-sided Fisher's exact test (corrected p-value = 1.237e-003) (for the null hypothesis that there is no enrichment of NHR in the considered 11 families as compared to all 22 AMPcg families). The correction factor was 440 that equals to the number of identified motifs in all 22 families. The parameters for p-value were: k = 92 (number of motifs that potentially bind NHRs in the group of families), n = 220 (number of motifs identified in 11 families), K = 139 (total number of motifs in all 22 families that potentially bind NHRs), N = 440 (total number of all identified motifs in all 22 families). The small p-value suggests that the enrichment of motifs that potentially bind NHRs is statistically highly significant for the considered 11 AMPcg families out of 22 analyzed families.

Based on distribution of absolute number of TFBSs in different categories it was observed that Spag11, an epididymis-specific defensin, which is also important in inducing sperm maturation (Zhou *et al.*, 2004) appears to be distinctly regulated compared to other members of the beta-defensin family (Yamaguchi *et al.*, 2002). The

data shows that spag11-specific motifs are over-represented, compared to motifs of the beta-defensin family (**Table 5.5**).

On the other hand, if looking at the rank position of a particular TF group in individual AMPcg families (number of motifs that could bind TFs from a particular category), six TF categories emerge as dominant categories. These are, in order, liver-specific, neuron system-specific, adipocyte-specific, NHR, immune-cell specific and lung-specific TFs. The ranking of TFs suggests that the functions of AMPs extend far beyond antimicrobial actions as mediators in energy metabolism and neuroendocrine regulations. The finding is reminiscent to multi-functionality of cytokines (i.e. IL6, TNF-alpha, MIF etc.) in adipocytes, liver and immune cells during metabolic challenges and stress (Mohamed-Ali *et al.*, 1998, Yudkin *et al.*, 2000 and Sakaue *et al.*, 1999). The results are presented in (**Supplementary Table** 5.5)

As a further support to the above finding, comparison with another dataset was performed to ascertain the claim that the TF groups that have been found to influence AMP gene groups are not non-specific for AMPs. To test this, motif search was carried out in similar manner on a set of 78 promoter sequences from non-immune, house keeping genes, considered as a negative data set. To determine if the usage of different categories of TFs is the same in AMP-genes and house keeping genes, a non-parametric ranksum test was performed (Conover, 1998). The ranksum test allows evaluation of actual population distributions rather than means of populations, which would be the typical test used to compare AMP families versus housekeeping genes.

Individual rank sum test was carried out for each of the six significant TF categories (AD, NHR, IMM, LIV, LUNG, NS), comparing the numbers of TFs in all

AMP families with the housekeeping gene set. The ranksum test gave a p-value is 5.1872e-004 and the corrected p-value for 6 tests was 0.0031, indicating that the null hypothesis can be rejected, that assumes the same population of TFs influence AMP genes and housekeeping genes. Consequently, it is concluded that AMP genes and housekeeping genes utilize different groups of TFs in a significantly different manner. Among the six most utilized TF categories (AD, NHR, IMM, LIV, LUNG, NS) NHR was found indicating a potential link of the endocrine and immune response systems (**Supplementary Table 5.6).**

Table 5.5: Distribution of motifs associated with different tissue/function-specific TF groups among AMP families.

Tissue/function-specific TF groups are AD: adipocyte-related TFs; NHR: nuclear hormone receptor TFs; CC: cell cycle-related TFs; IMM: immune cell-specific TFs; LIV: liver cell-specific TFs; LUNG: lung cell-specific TFs; MUS: muscle cell-specific TFs; NS: nervous system-related TFs; PAN: pancreatic B-cell related; PIT: pituitary gland-specific TFs; Eye: eye-specific TFs; BS: bone-specific TFs. TF groups (AD, NHR etc.) that occur with highest frequency among AMP families are underlined. Cut-off indicates the minimum percentage of motifs in a TF family that can bind TFs from a particular tissue/function-specific group.

Tissue/function-specific TF groups	AD	NHR	CC	IMM	LIV	LUNG	MUS	NS	PAN	PIT	EYE	BS	Cut-Off (%)
Total no. of motifs	131	139	97	122	141	122	78	143	77	74	1	12	
No. of AMPcg families	17	18	8	14	19	16	4	17	5	5	0	0	25%
	14	12	7	11	14	10	4	15	3	2	0	0	30%
	8	11	3	6	10	5	2	11	2	1	0	0	35%
	4	7	3	4	3	4	1	9	0	0	0	0	40%

AMPcg Families	AD	NHR	CC	IMM	LIV	LUNG	MUS	NS	PAN	PIT	EYE	BS
Alphadefensin	6	12	2	6	6	5	3	9	2	1	0	0
Apoa2	5	5	5	4	5	5	4	6	4	3	0	0
Betadefensin	6	5	4	6	6	5	2	8	2	3	0	1
bin1b/spag11	9	9	3	5	10	10	6	10	5	2	0	3
Bpi	6	7	8	8	8	5	8	7	7	4	0	0
Calgranulin	8	5	6	9	9	6	7	11	7	5	1	2

Cathelicidin	4	8	6	8	5	5	1	6	3	5	0	1
Dbi	7	8	4	5	7	6	4	6	1	4	0	0
Slpi	6	7	3	6	6	5	3	8	5	4	0	0
Granulin	6	5	6	4	6	6	4	5	4	3	0	0
Hepcidin	10	9	3	7	11	11	3	9	6	7	0	0
Histone	5	2	3	3	5	4	3	3	4	3	0	0
Lactoferrin	7	10	3	4	7	6	4	8	3	1	0	0
Lysozyme	4	2	4	3	4	4	3	3	2	1	0	0
Mbp	6	7	9	9	7	6	6	7	2	6	0	2
Melanotropinalpha	9	6	9	7	8	7	4	8	3	4	0	0
Proenkaphalin	7	7	3	4	7	7	1	8	3	3	0	1
Secretogranin	1	5	2	6	3	2	4	3	3	3	0	1
Spyy	5	5	2	5	5	3	1	5	3	5	0	0
Vip	3	3	3	4	3	2	1	3	3	1	0	1
Vstn	4	4	3	3	5	4	3	4	1	2	0	0
Zap	7	8	6	6	8	8	3	6	4	4	0	0

5.4.6 Motifs associated with nuclear hormone receptors (NHRs)

NHR proteins function as dimeric molecules in the nucleus to regulate the transcription of target genes in a ligand-responsive manner (Nishikawa *et al.*, 1995 and De Vos *et al.*, 1994). A number of PEs detected in the 22 AMPcg families potentially binds different TFs. Among them the most frequent is the family of NHRs that in the case includes (**Table 5.6**) AR (androgen receptor), GR (glucocorticoid receptor), RXR-alpha (retinoid X receptor alpha), VDR (vitamin D receptor), T3R-alpha (thyroid hormone receptor) and ER-alpha (estrogen receptor alpha), ERRalpha1 (Estrogen-related receptor alpha 1), RAR(retinoic acid receptor)-alpha, beta, gamma, LXR (liver X receptor)-alpha, beta, PPAR (peroxisome proliferator-activated receptor)-alpha, beta, gamma. GR, RXR-alpha, AR, VDR, T3R-alpha and RAR-alpha1.

Table 5.6: Distribution of individual TFs among AMP families

TF	No. AMP families with detected TF	AMP family names
GR	20	alpha defensin, apoa2, betadefensin, bin1b, bpi, calgranulin, cathelicidin, dbi, slpi, granulin, hepcidin, histone, lactoferrin, lysozyme, mbp, melanotropinalpha, penk1, vip, vasostatin, zap
RXR-alpha	18	alpha defensin, betadefensin, bpi, calgranulin, cathelicidin, dbi, slpi, granulin, hepcidin, histone, lactoferrin, mbp, melanotropinalpha, penk1, secretogranin, spyy, vip, zap
AR	17	alpha defensin, apoa2, betadefensin, bin1b, bpi, calgranulin, cathelicidin, dbI, slpi, granulin, hepcidin, lactoferrin, mbp, melanotropinalpha, penk1, vasostatin, zap
Sp1	16	apoa2, bpi, calgranulin, dbi, granulin, hepcidin, histone, lactoferrin, lysozyme, mbp, melanotropinalpha, penk1, secretogranin, spyy, vasostatin, zap
VDR	16	alpha defensin, apoa2, betadefensin, bpi, calgranulin, cathelicidin, dbi, slpi, granulin, hepcidin, lactoferrin, mbp, secretogranin, spyy, vasostatin, zap
T3R-alpha	15	alpha defensin, apoa2, betadefensin, bin1b, bpi, calgranulin, cathelicidin, dbi, slpi, hepcidin, mbp, melanotropinalpha, penk1, spyy, zap
Meis-1a	15	alpha defensin, apoa2, betadefensin, calgranulin, cathelicidin, dbi, slpi, granulin, histone, lactoferrin, lysozyme, mbp, secretogranin, spyy, vip
Meis-1b	15	alpha defensin, apoa2, betadefensin, bin1b, calgranulin, cathelicidin, dbi, slpi, granulin, histone, lysozyme, mbp, secretogranin, spyy, vip
RAR-alpha1	14	alpha defensin, apoa2, betadefensin, bpi, calgranulin, cathelicidin, dbi, slpi, hepcidin, mbp, penk1, secretogranin, spyy, zap
LXR-alpha:RXR-alpha	13	alpha defensin, apoa2, betadefensin, bpi, calgranulin, cathelicidin, dbi, slpi, lactoferrin, mbp, secretogranin, vip, zap
NF-1	13	apoa2, bin1b, calgranulin, cathelicidin, dbi, granulin, histone, lactoferrin, lysozyme, mbp, melanotropinalpha, vip, zap
AP-2alphaA	13	apoa2, bin1b, bpi, cathelicidin, slpi, granulin, hepcidin, lysozyme, mbp, melanotropinalpha, penk1, spyy, vasostatin
Nkx2-1	12	betadefensin, bin1b, bpi, calgranulin, cathelicidin, dbi, slpi, granulin, lysozyme, spyy, vip, zap
c-Myb	12	bpi, calgranulin, cathelicidin, dbi, slpi, granulin, lysozyme, mbp, melanotropinalpha, vip, vasostatin, zap

Table 5.6: Comments

GR is involved in the regulation of numerous physiological processes including lymphocyte e apoptosis, T cell development and inflammatory responses (Reichardt, 2004). Several of the TFs found in the analysis are known to interact with GR, like AP-1, c-Ets-2 etc. AR has also been shown to play a role in the immune response. It appears that androgens have an influence on the developmental maturation of T and B lymphocytes (Olsen and Kovacs, 2001, Takeuchi *et al.*, 1998). RXR-alpha binds to many other TFs forming complexes that can regulate multiple pathways, including immunomodulatory pathways. It has been shown that RXR-alpha binds to VDR, forming a heterodimer that inhibits NF-AT and plays a role in immunosuppression (Takeuchi *et al.*, 1998). RXR-alpha also binds to PPAR-gamma and causes an apoptotic signaling cascade in B cells through NF-kappaB activation (Schlezinger *et al.*, 2002).

VDR is the receptor protein for 1,25-dihydroxyvitamin D which is involved in regulating cell growth, modulating the immune system and the renin-angiotensin system (Holick, 2003). Recently, it has been shown that VDR can mediate the induction of antimicrobial peptide gene expression in human like beta-defensin 2 (Wang T.T. *et al.*, 2004, Wang Y. *et al.*, 2004). The analysis of the beta-defensin family also shows the presence of VDR. T3R-alpha is another of the TFs to be found in high occurrence, covering 15 of the AMPcg families. It has been shown that T3R-alpha binds to thyroid hormone and is involved in the control of B-cell production level (Arpin *et al.*, 2000). RAR-alpha1 is a receptor for retinoids and it is constitutively produced in adenoidal T and B cells (Ballow *et al.*, 2003). LXR-alpha: RXR-alpha heterodimers function as sensors for

cellular oxysterols and, are transcriptional activators of genes that control sterol and fatty acid metabolism/homeostasis (Edwards *et al.*, 2002). In summary, the occurrence of different families of NHR as most frequently occurring TFs among AMPcg families indicates an intricate regulatory network encompassing the endocrine (i.e. lipid metabolism) system and innate immunity system (**Table 5.6**).

NF-1 (nuclear factor 1) is known to be involved in regulation of genes associated with adipogenesis and signal transduction pathways induced by steroid hormones like vitamin D, thyrotropin Gronostajski, 2000. The AMP member diazepam binding inhibitor (Dbi), is known to have an NF-1 site that plays a crucial role in its transcription in the lipogenesis pathway (Hansen *et al.*, 1991).

5.4.7 Other TFs and their potential role in AMPs

Several non-NHR TFs that frequently appear in genes of the 22 AMPcg families were also found. (**Table 5.6**). Sp1 is a ubiquitous TF that is enriched in the numerous GC-rich housekeeping gene promoters, but also contributes to tissue-specific transcription. For example, it is detected in the promoters and enhancers of numerous erythroid cell-expressed genes and appear to cooperate with lineage-restricted factors in directing their expression (Suico *et al.*, 2004). Meis1a and Meis1b isoforms are homeoproteins related to the pre-B cell transformation protein family. Meis1a is implicated in the myelopoesis (Calvo *et al.*, 2001) leading to the basophil, neutrophil and eosinophil granulocytes. Meis1a and Meis1b binding sites were detected in members of the apoa2, calgranulin, slpi, granulin, secretogranin, mbp, vip, lysozyme AMP families, suggesting a granulocyte-specific transcriptional control function. Calvo and co-workers (Calvo *et al.*, 2001) showed that Meis1a suppressed the G-CSF-induced transcription of neutrophil

differentiation-specific genes cytochrome b-245 beta, lactoferrin, early growth response-1, neutrophil gelatinase B, and lipopolysaccharide receptor CD14. The unique C-terminus of Meis1a which was shown to specifically mediate protein kinase A and trichostatin activation (Huang *et al.*, 2005) provides additional support for the functional differences of Meis1a and Meis1b. Meis1a in combination with other neutrophil-specific TFs (i.e. STAT1, STAT6 and NF-kappa B) may play an important role in the recruitment and activation of neutrophils seen in sepsis and Helicobacter pylori infection-induced iron-deficiency (Baveye *et al.*, 1999, Choe *et al.*, 2003). Interestingly, hepcidin, which inhibits iron absorption from the small intestine during infection-induced inflammation, lacks Meis1, suggesting the induction of multiple alternative transcriptional regulation mechanisms during microbial pathogenesis.

5.4.8 Suggested future experiments

The analysis has generated a number of hypotheses that are in good concordance with some of the existing knowledge in the field. However, the computationally-inferred hypotheses can only be validated through experiments. The author proposes the following hypotheses which warrants for experimental validation.

1. NHRs maybe involved as dominant group in regulating AMP genes. NHR candidates such as GR, RXR-alpha, AR, T3R-alpha, RAR-alpha, LXR-alpha:RXR-alpha should be tested for their presence in the promoter regions of AMP genes.

2. VDR which is already known to be involved in directly regulating expression of beta defensins, also maybe involved in regulation of many other AMP genes as

listed in Table 5.6. Hence, the presence of VDR binding site should be validated experimentally in other AMP genes.

3. NF-1 and NKX2-1 which have not yet been implicated to be involved immunomodulatory pathways have appeared frequently in many AMP genes in the analysis.

4. *C-myb* transcriptional regulator is known to be invoved in cell proliferation, differentiation (Farrar *et al.*, 1989, Ramsay, 2005). It is critical in lymphocyte development (Thomas *et al*., 2005). A hypothesis related to its role in neuroectodermal tumors alludes to activation of innate immune pathway due to inhibition of *c-myb* by antisense oligodeoxynucleotides (Pastorino *et al*., 2004*)*. However, there is no consolidated evidence of its involvement in innate immunity.

5. Meis1a as discuseed in the previous section has been implicated in regulation of lactoferrin, may also be involved in regulation of other AMPs under certain conditions.

The author proposes using microarray technology combined with chromatin immunoprecipitation (ChIP) profiling (Ren *et al.*, 2000) to identify all the chromosomal locations that are occupied by a transcription factor. These experiments are expected to clarify which promoters and TFs are specific for certain tissue cells and how many AMPcgs are regulated by a TF, TF pair or multiple TFs. Eventually, the combination of both computational and experimental should permit us to construct mechanistic models of AMPcg regulatory transcription networks.

5.5 Conclusion

The large-scale computational analysis of promoters of 22 families of AMPcgs across three mammalian species has allowed us to identify potential key transcription elements of these families. Promoter regions (-1000, +200) were analyzed and it is likely that the regulatory elements further upstream may have been missed, that might be important in the fine-tuning of the regulation of particular families of AMPcg. The results suggest a core set of transcription factors (TFs) that regulate the transcription in the mouse, rat and human AMPcg families examined. TFs of the liver, nervous system- specific and NHR group are significantly over represented. These TF groups consist of transcription regulators that are involved in diverse physiological functions, including control of embryonic development, cell differentiation and homeostasis, and also in immune response. Interestingly, NHRs appear more dominant than immune cell-specific TFs in the analyzed AMPcg families. Numerous experimental evidence show the involvement of NHRs in various immunomodulatory pathways (Reichardt, 2004, Hayes *et al.*, 2003, Jeay *et al.*, 2002, Reichardt *et al.*, 2000). However, little is known about their direct involvement in innate immunity. Recently, there has been evidence that VDR plays a direct role in the induction of antimicrobial innate immune response (Wang T.T. *et al.*, 2004). This analysis concurs with this evidence and elucidates other members of the NHR family also, that could play a crucial role in antimicrobial innate immunity. Besides the NHR, putative binding sites of Sp1, Meis1a, Meis1b, NF-1 were also found to be prevalent across different AMPcg families. Three potential TF-binding motifs that are enriched in promoters of AMPcgs are novel. Four identified motifs were found to be species-specific. Phylogenetic analysis of alpha-defensins revealed potential TF-binding

motifs and motif combinations that are common in primates and rodents, and others that are species-specific and specific to enteric versus myeloid expression of alpha-defensins.

This analysis brings out the advantage of using a computational approach to analyze promoter regions, since the author was able to do a comprehensive analysis and get a bird's eye view of the transcriptional regulators involved in multiple AMPcg families across different mammalian species.

In addition, 102 new motifs were discovered as candidate TFBS with a role in antimicrobial innate immunity. The actual experimental confirmation of the AMPcg transcription regulatory elements can only be accomplished by targeted research of infection or cellular stress models using time-course sampled tissue cell types.

After finding potential TFBS motifs for several of the AMP gene groups within different AMP families, it intrigues to know which of the TFBS motifs can appear together across promoter regions of same AMP gene across different species. It is probable that co-occuring TFBS motifs that are conserved across different species for a gene or conserved across a gene family have a role to play in transcription regulation and hence are not present in the regulatory region by chance. Chapter 6 elucidates this hypothesis in detail.

Part III: Chapter 6 Identification of transcription factor binding site modules

The power of imagination makes us infinite.
(John Muir)

6.1 Introduction

In this chapter, the author analyzes in greater detail promoter regions of three AMP gene groups (Alpha defensin, Penk, Zap) that appeared in Chapter 5 so as to identify transcription factor (TF) binding motifs that are common among AMP genes of mammalian species (i.e. namely human, mouse and rat). In the case of alpha-defensins and penk, experimentally identified promoter elements were used to assess and interpret the predictions. For the zap family, the findings are novel, since no experimental data is present. Further, the author has attempted to identify Transcription Factor Binding Site modules (TFBS) module(s) or promoter models which are defined as a TF framework consisting of more than one motif found within a given distance and orientation (Werner *et al.*, 2003).

Identification of TFs that control the expression of a given gene is the first step to towards understanding the transcriptional regulatory network associated with a gene or a given class of genes. TFs mediate their effects via their cognate TFBSs. TFs work in combinations to bring out special-temporal expression of genes. Thus, a set of TFs that modulate a functional response may trigger a set of related genes associated with that functional response. Therefore, finding TFBS modules can lead us to predict other genes that are responsive to the same set of TFs (Dohr *et al.*, 2005).

A **TFBS module** or framework is a model consisting of two or more TFBSs found within a certain distance, having a defined order relative to each other and having the same strand orientation. It has been shown that TFBS organization in the promoter region plays an important role in transcriptional regulation (Fessele *et al.*, 2001). Genes

expressed in the same tissue under similar conditions often share a common organization of regulatory binding elements. This organization appears to be conserved across different species whereas structure and function of a gene product may be more tolerant of gene mutations in the coding sequences. The specific arrangement of TFBSs increases the potential specificity of the system to affect gene mechanisms as co-regulation imposes stringent constraints on the evolution of the gene's promoters. Thus the organization of promoter motifs can give essential clues about the transcriptional regulatory mechanisms at work in a specific biologic context and provide information about signal and tissue specific control of expression (Werner *et al.*, 2003). It can thus be considered as a "footprint" or "signature" of transcriptional regulatory mechanisms at work in a specific biologic context (Werner *et al.*, 2003). **Figure 6.1** shows an example of a TFBS module.

Dushay *et al* showed (Dushay *et al.*, 2000) that promoter region of Drosophila AMPs, cecropin, diptericin, metchnikowin, attacin A and attacin B have a common TFBS module (Werner *et al.*, 2003). The proximity of GATA, R1 and ICRE to kappaB sites was shown (Dushay *et al.*, 2000) to be important for gene expression, as removal of these sites reduced the expression of cecropin and diptericin, despite the presence of intact kappaB sites. Thus, the presence of certain TFBSs in a particular order and position indicates a certain way of conservation that could be essential for induction of a particular gene. It is not just a random occurrence of these motifs in that region. Moreover, since these subregions of conserved positional arrangement of promoter motifs are usually sited at various distances from TSS in different species, they cannot be easily detected in most cases by the usual local alignment methods.

6.2 Background

Prior studies have suggested that promoter modules are pathway-specific or cell- type-specific and hence cause the transcriptional response to specific signal transduction pathways, cell type-specific expression and events central to developmental regulation.

A study (Werner *et al.*, 2003) done of RANTES/CCL5 gene set corroborates this point. RANTES/CCL5 is a member of the -CC- subfamily of chemotactic cytokines which is involved in different stimuli and plays diverse roles in inflammatory processes. Analysis of the RANTES promoter in different cells like monocytes, T cells, astrocytes and mesangial cells show that there is an underlying group of six functionally characterized short regulatory elements forming a hierarchic organization in them. However, the combinations of these elements vary in the four different cell types (Werner *et al.*, 2003). This sequence feature can be exploited to look for genes that are regulated by similar mechanisms.

To predict TFBS module, gene expression (Segal *et al.*, 2003, Ihmels *et al.*, 2004, Kloster *et al.*, 2005, Kloster *et al.*, 2005, Wang *et al.*, 2005) or DNAseI hypersensitivity data (Noble *et al.*, 2005) are taken into account. Most of the methods derive a set of TFBS elements from a set of co-regulated genes or a set of genes with similar functions. The individual binding elements are then combined into one recognition module/model.

Another approach is that a set of transcription-factor PWMs that are known to be co-occuring are used to identify genomic regions densely populated in putative sites for these TFs (Bailey and Noble, 2003, Frith *et al.*, 2003, Johansson *et al.*, 2003, Sinha *et al.*, 2003, Alkema *et al.*, 2004).

To find functional binding sites and modules, the concept of phylogenetic footprinting is also used. Phylogenetic footprinting is a comparative genomic approach by which non-coding regions of orthologous genes from different species that are sufficiently evolutionarily distant (but not too distant) are aligned to detect the conserved regulatory elements interspersed between the real non-functional background sequences (Zhang and Gerstein, 2003). The major advantage of phylogenetic footprinting compared to other techniques is that it is capable of identifying regulatory elements specific even to single genes, as long as they are sufficiently conserved across species. This approach facilitates finding functional binding sites. However, it is important to note that many of the TFBSs that are detected computationally in the promoter region may not be functional. They may be false positives, or actually binding sites that are not used in the context of the gene studied. By comparing these sequences across species, phylogenetic footprinting can help reduce this problem to an extent. **Table 6.1** lists the various transcription factor module finding programs.

Table 6.1: Transcription factor module finding programs

** PMID – Pubmed Unique Identifier*

Program name	URL Address	Reference (PMID)*
MSCAN	http://mscan.cgb.ki.se/cgi-bin/MSCAN	15215379
MAST	http://meme.sdsc.edu/meme/mast.html	9520501
Cluster Buster	http://zlab.bu.edu/cluster-buster/cbust.html	12824389
CRÈME	http://creme.dcode.org/	12855471
Module Scanner		14534164
Dragon Promoter Mapper	http://defiant.i2r.a-star.edu.sg/projects/BayesPromoter/	16613910
MCAST		14534166

6.3 Materials and methods

Figure 6.1: Graphical representation of TFBS module generation

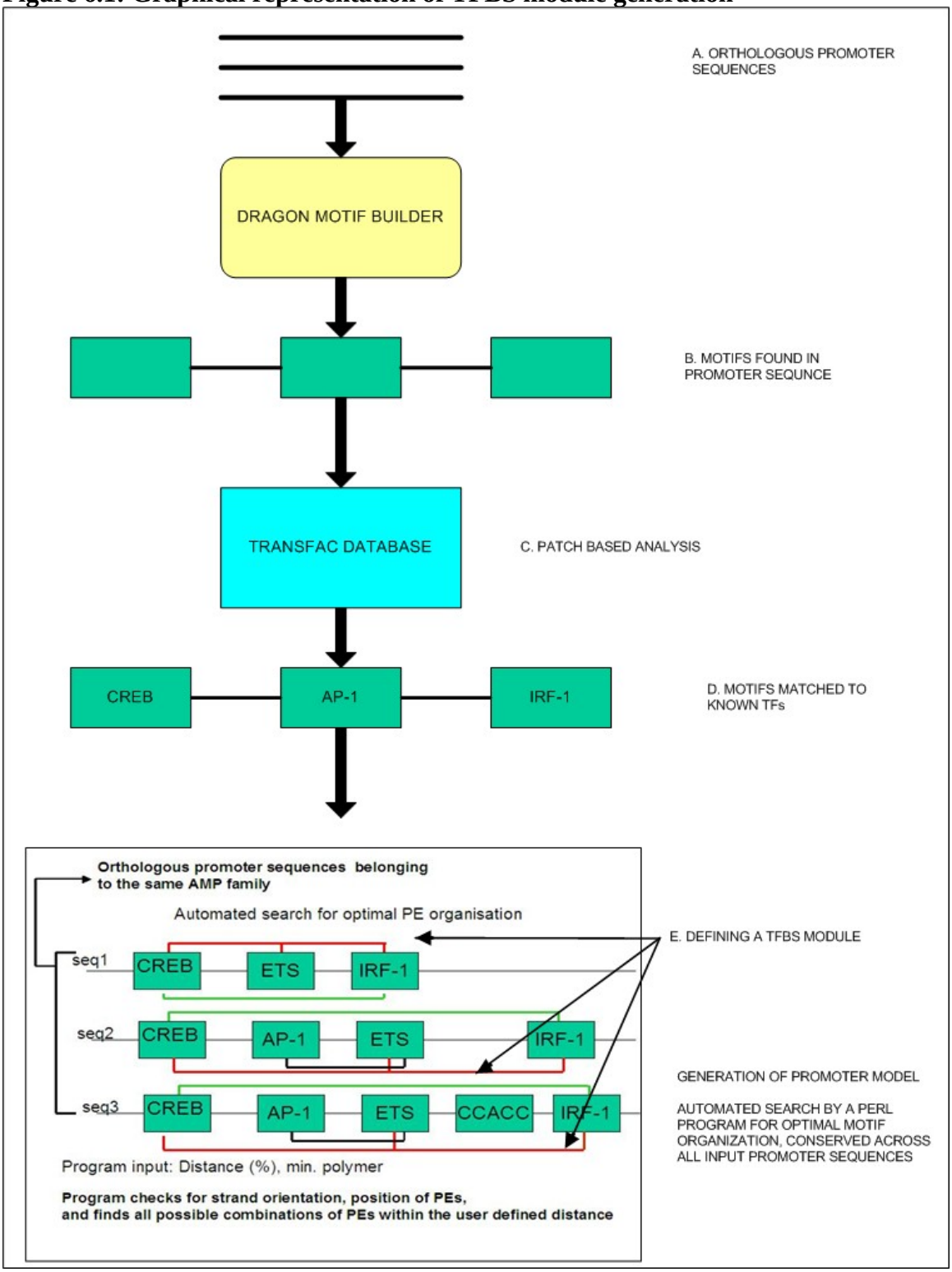


Figure 6.1 : (A,B): Orthologous promoter sequences are subjected to DMB motif finding program. (C,D): The motifs found by DMB are searched in the TRANSFAC database for known TFBSs using Patch program (TRANSFAC program). E: TFBS modules are generated by looking for all possible combinations of motifs common across all the input sequences which have the same relative order and strand orientation. The different colored line connectors (red,green,black) highlight the three different TFBS modules found by the program within a given distance cut-off.
The possible combinations shown here are CREB-ETS-IRF-1, AP-1-ETS,CREB-IRF-1. In this study the minimum number of motifs (min.polymer) is set to 3 or more, therefore combinations below the min.polymer threshold will not be shown in the final output by the program. PE: promoter element

6.3.1 Data selection for generation of promoter models

Alpha-defensins DEFA5, DEFA1, Penk and Zap promoter sequences covering (-1000, +200) relative to the estimated transcription start site (TSS) from human, mouse and rat were selected. The method of extraction was same as discussed in Chapter 5. Promoters of human orthologs were extracted using H-invitational database (Fujii *et al.*, 2004) as well as PromoSer (http://biowulf.bu.edu/zlab/PromoSer) (Halees *et al.*, 2003s). All these resources provide estimated TSS locations based on mapping EST and full length cDNAs (flcDNA) data to genomic sequences. The TSS location of mouse sequences was determined by using the start position of the first exon of the FANTOM cDNA-genome mapping data (http://fantom31p.gsc.riken.jp/cage/download/mm5/cage.rep_tag.2004-11-16.chr_all_gff.tar.gz). Mouse promoter sequences (-1000, +200) were then extracted by mapping the TSS location to the mouse genome data from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/mm5/chromosomes/).

6.3.2 Prediction of TFBS motifs

The author aligned the (-1000,+200) promoter regions of the orthologs of an AMP family and subjected them to Dragon Motif Builder (DMB) (E Huang *et al.*, 2005) program for prediction of TFBSs. The EM (Expectation Maximization) threshold of the DMB program was set to 0.85 for all AMP groups except for DEFA1, where it was set to 0.70 as 0.85 appeared too stringent. The number of 10-15 nt motif candidates was restricted to a maximum of 20 for each of the AMPcg families. The author chose to search for maximum of 20 motifs as the optimum number that covers most of the promoter region of (-1000, +200) without overlap of motif candidate sequences. After DMB identified the sequence motifs, the Patch program (mismatch =0; motif length =6; species =all) (Wingender *et al.*, 2000) of the TRANSFAC database ver. 8.4 was used to infer potential transcription factors (TFs) that may bind to these motif families. The motifs are reported in IUPAC nucleic acid codes format.**Figure 6.1** shows the schema for the generation of TFBS modules.

6.3.3 Generating the TFBS models

The author searched for all possible combinations of motifs that were present in same order and same strand orientation for a given set of promoter sequences within a family and constrained by a defined range of distances between the motifs. A perl program (made by I2R Knowledge Discovery Group) was used that calculated TFBS models from the graphic motif representation file generated by DMB and user input of distance constraint in percentage and minimum number of motifs. Promoter models that contained

experimentally proven TFBSs – if available – for a particular AMP gene group along with other motif candidates or TFBSs within the given distance constraint, were selected. The minimum number of motifs per model was set to three. The distance constraint was tested for the interval of 1%–30% and promoter sequence length of 1200 bp. It was observed that promoter models having three or more motifs could be generated with distance percentages of 20 or 30. This distance percentage appeared to be optimal for the length of 1200 bp. Hence, the distance between two adjacent motifs in a promoter model ranges between one to 240bp or up to 300bp from each other. The motif combinations that appeared common across all promoters of a given AMP family were chosen as candidates for scanning the large promoter data set.

6.4 Results

6.4.1 Alpha defensin promoter model

Human polymorphonuclear leukocytes (PMNs) or neutrophils express four defensins named human neutrophil peptides, HNP1 to HNP4. HNPs are also expressed in immature bone marrow cells, in HL-60 and human promyelocytic leukemia cells. Alpha defensin 5 are enteric defensins expressed mainly in the Paneth cells and are constitutively produced (Cunliffe, 2003). Gene duplication events have probably led to both species-specific and functionally diverse subsets of alpha-defensins,which should be also reflected in the upstream regulatory regions. For example, enteric-expressed defensins are important to barrier function of the gut mucosal surface against bacteria, whereas myeloid and

neutrophil-specific defensins help macrophages and neutrophils to kill internalized bacteria (Patil *et al.*, 2004).

The author analyzed HNP-1 (DEFA1) and HNP-3 (DEFA3) from human, which are paralogs and alpha-defensin 5 group orthologs (information extracted from Ensembl release 40- Aug 2006, http://www.ensembl.org). DEFA1 and DEFA3 have very similar promoter regions and are controlled by myeloid-specific regulation, even though they have different biochemical properties (Tsutsumi-Ishii *et al.*, 2000).

The promoter model of the alpha-defensin 5 group contained four motif candidates in the order 15-12-19-5 within the distance threshold of 360 nt (**Table 6.2**, **Figure 6.2a**). Motif 5 and motif 15 did not correspond to any known TFBS. Motif 12 represented potential binding sites for TFs namely LF-A1, GATA-1, COUP-TF2, NKX2-1, NF-E3. GATA-1 is up-regulated by cytokine IL-1B during inflammatory responses (Chuen *et al.*, 2004). Recently, it was reported that human alpha defensin 5 (HD-5) binds to the cell membrane of intestinal epithelial cells and induces secretion of the interleukin (IL)-8. HD-5 may be playing a role in regulation of the intestinal inflammatory response (de Leeuw *et al.*, 2007). Hence, the presence of putative GATA-1 binding site in DEFA5 may indicate binding of GATA-1 to its promoter region and causing up-regulation of its expression during inflammatory response as a positive feedback loop. Motif 19 corresponded to potential binding site for C\EBPalpha. Transcription factor C/EBPalpha is known to bind to promoter region of HNPs and regulate myeloid-specific genes (Tsutsumi-Ishii *et al.*, 2000).

The analysis of the promoter regions of DEFA1 and DEFA3 genes returned motifs for putative binding sites for CCAAT-binding factor, NF-Y represented by motif

10. Zic2 was represented by motif 20. Motif 14 corresponded to binding site for Ets transcription factor (c-Ets-2) binding site. Ets is known to bind to HNP1-3 promoter region (Tsutsumi-Ishii *et al.*, 2000). Motif 5 represented putative binding sites for HNF-4, HNF-4alpha1, C/EBPalpha, C/EBPbeta. As the promoter regions for DEFA1 and DEFA3 are highly conserved, the original model showed that all the 20 motifs have a conserved organization across the two promoter regions. Thus, a subset of consecutively positioned motifs was selected and a promoter model created that contained the motifs 10-20-13-5 within a maximum allowable distance range of 240 nts between them (**Table 6.2, Figure 6.2b**).

Table 6.2: Alpha defensin promoter models

Hs: *Homo Sapiens*, Mm: *Mus musculus*,

AMP group	Motif No.	Motif	Species	Start	End	Putative TFBS	Distance range
alpha defensin 1	10	TTAGCCACAGCCAAT	Hs	737	751	CCAAT-binding factor (CTF CTF-1 CTF-2 CTF-3 NF-1) NF-Y	240
			Hs	730	744		
	20	AGTTGGTTGCTGCCT	Hs	794	808	Zic2	
			Hs	787	801		
	14	CCTTCCCACCAAATT	Hs	873	887	c-Ets-2	
			Hs	866	880		
	5	ATGGACCCAACAGAA	Hs	919	933	HNF-4 HNF-4alpha1 C/EBPalpha C/EBPbeta	
			Hs	912	926		
alpha defensin 5	15	GAAKMCTGCAR	Mm	19	29	unknown	360
			Hs	582	592		
	12	YMACACMTTGGRYY	Mm	223	236	LF-A1, GATA-1, COUP-TF2, NKX2-1, NF-E3	
			Hs	799	812		
	19	RGAGGSATKRA	Mm	487	497	unknown	
			Hs	817	827		
	5	YATCCTTGCTG	Mm	871	881	C\EBPalpha	
			Hs	1057	1067		

Figure 6.2a: Motif arrangement in promoter region of mouse Defcr3 and its human ortholog (DEFA5)
The numbers (i.e. 15) in the boxes refer to different motifs. The grey line represents the (-1000, +200) promoter region that has been analyzed for motifs. The broken arrow indicates the TSS.

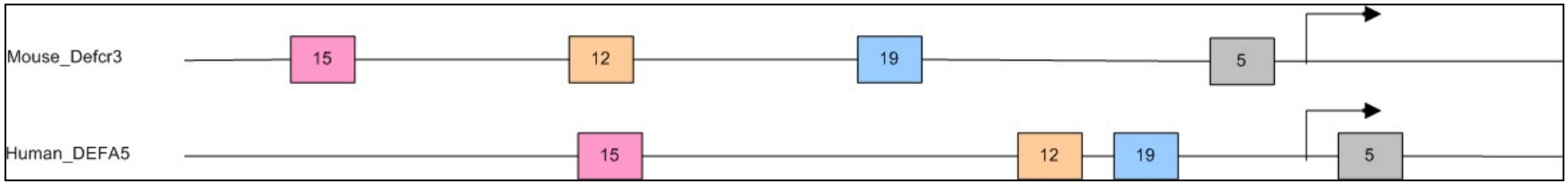
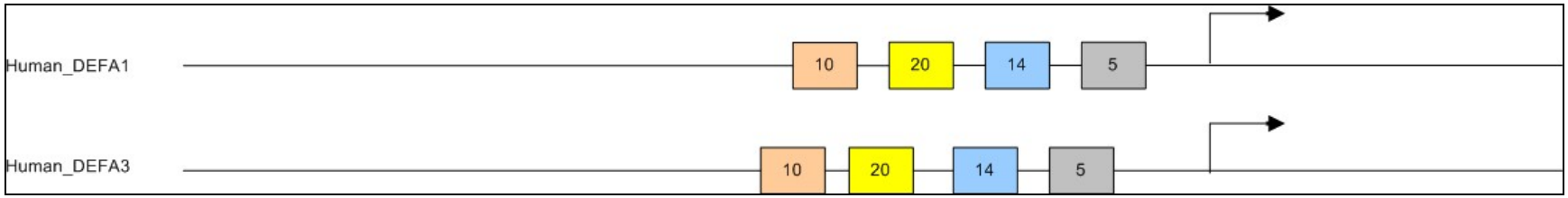


Figure 6.2b: Motif arrangement in promoter region of human DEFA1 and its human paralog DEFA3
The numbers (i.e. 10) in the boxes refer to different motifs. The grey line represents the (-1000, +200) promoter region that has been analyzed for motifs. The broken arrow indicates the TSS



Both the alpha defensin promoter models have TFBSs such as C/EBP and Ets that have experimentally proven to be present in the regulatory regions of this group of AMPs. This corroborates the viability of the strategy that has been implemented in this thesis to find TF motifs and promoter models. The other new TF motifs such as Zic2, NF-Y etc. could be interesting candidates for experimental validation in alpha defensin promoter regions.

6.4.2 Promoter models of penk and zap families

Penk and zap gene groups were chosen for detailed promoter analysis because penk1 represents a gene that has several experimentally identified promoter elements and this could be used as a benchmark to assess the accuracy of the computational analysis. For the zap family, to the best of the author's knowledge, the results presented are completely new.

6.4.3 Penk promoter analysis

Penk1 is a neuropeptide-encoding gene that is known to be expressed primarily in mature nervous and neuroendocrine systems. The penk1 gene product is known to mimic the effects of opiate drugs. It plays a role in a number of physiologic functions, including pain perception and responses to stress. Experimental evidence shows that penk family members are also expressed in activated lymphocytes (Ovadia *et al.*, 1996). Their expression is induced by bacterial endotoxins (i.e. lipopolysaccharide, LPS). Penk-derived peptides have immunomodulatory

properties ranging from augmenting CTL and NK cell, monocyte chemotaxis to being involved in pathophysiology of endotoxic shock (Ovadia *et al.*, 1996, Salzet, 2001). The penk1 promoter regions contain the experimentally characterized motifs for AP-1, CRE, NF-1, AP2, NF-Y, NF-kappaB, MZF-1 and PACH-1 (Le *et al.*, 2003, Liu *et al.*, 2000). The computational analysis was able to identify all except PACH-1 because TRANSFAC 8.4 database lacked the corresponding motif. The motifs and corresponding TFBSs associated with the penk family are listed in **Supplementary Table 6.1.**

In different enkephalin-expressing tumor cell-lines, as well as in adult enkephalinergic neurons, the rate of transcription of penk is modulated by several ubiquitous factors like NF-kappaB, AP-2, cAMP-response element binding protein, etc. (Le *et al.*, 2003), whose DNA binding sites are located immediately upstream of TSS of penk (Uhl *et al.*, 1991). This 200 bp DNA stretch is extremely well conserved among human (Comb *et al.*, 1992), rat (Joshi and Sabol, 1991) and mouse (v Agoston *et al.*, 1998) promoter regions.

The computational analysis showed that motif 2 corresponds to the TATA box. Motif 5 represents a potential binding site for NF-kappaB and AP-2. Motif 9 contains the putative binding site for NFI/CTF, and NF-Y. Motif 16 represents potential binding sites for MZF1, AP2 and NF-kappaB. Motif 3 corresponds to GR and AR. It has been found previously that GR is involved in activation of cAMP-mediated transcription of penk in rats (Jenab and Inturrisi, 1995). Motif 1 appears as NHR representing motif. It contains the potential binding site for RXR-alpha, LXR-alpha, ERRalpha1. An AP-1 site is known to be present in the penk promoter

(Macian *et al.*, 2001) and is represented by motif 6. Motif 7 represents binding sites of c-Ets1, Elk-1, SAP-1a, SAP-1b, PEA3 and ELF-1, all of which belong to the Ets family of transcription factors. The expression of penk gene in epididymis is regulated by testicular factors that control expression via members of the Ets transcription family, (Hinton *et al.*, 1998). Motif 12 represents putative USF family of transcription factors USF-1, USF1, USF2, USF2b and USF. These TFs are involved in the regulation of activity-dependent gene expression in neurons (Chen *et al.*, 2003). They are found along with CREB (cAMP response element binding) binding elements in a number of promoters (Cvekl *et al.*, 1994, Durham *et al.*, 1997, Kingsley-Kallesen *et al.*, 1999) which indicates that the these two factors cooperatively activate transcription of calcium-inducible neuronal genes. Moreover, the rat penk gene is also known to be regulated by cyclic AMP and calcium pathways (Konradi *et al.*, 2003), supporting the observation of USF TFBSs in the promoter of proenkephalin genes (**Supplementary Table 6.1**)

6.4.3.1 Penk family promoter model

The motifs shown in **Table 6.3** were analyzed in terms of their orientation, positioning and mutual distance common to all three sequences (mouse, human, rat) to create promoter models. The arrangements of the motifs are shown in (**Table 6.3** and **Figure 6.3**). A single model arrangement 3-5-1-13 (representing motifs 3, 5, 1, 13) common to all the three considered species was generated. The corresponding TFs that may bind to the motifs are: GR, AR (motif3), NF-kappaB, AP-2 (motif5). Motif1 represents potential binding sites for NHR (RXR-alpha, LXR-alpha,

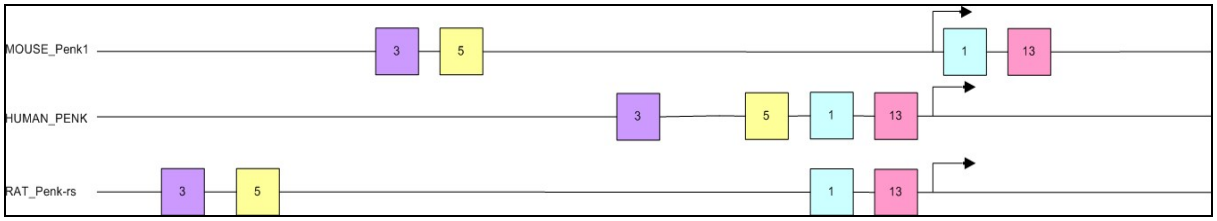
ERRalpha1), while Motif13 potentially binds to various TFs including NHR (DSF GCN4 COUP-TF1 RAR-beta RXR-alpha RAR-alpha1 TLX Pax-2.1).

Table 6.3: Motif arrangements in promoter region in mouse (4922504O09), human (HIX0007519.2) and rat (NM_017139) of Penk family members.

Species	Motif arrangement
Hs,Rn	3-5-15-18-4-16-8-9-2-10-1-13
Mm, Hs	7-12-3-5-1-13
Hs, Rn	3-5-1-13-20
Mm, Hs, Rn	3-5-1-13

The species abbreviations are Hs: *Homo sapiens*; Mm: *Mus musculus*: Rn: *Rattus norvegicus*.

Figure 6.3 Conserved Penk motif organization in mouse, rat and human



The numbers (i.e. 1) in the boxes refer to different motifs. The grey line represents the analyzed (-1000, +200) promoter sequence. This figure is a graphical representation of the common motifs across sequences determined by DMB program. It is not drawn to scale. The broken arrow indicates the TSS.

6.4.4

6.4.4 Promoter elements and their organization in the zap family

The AMP CCCH-type zinc finger protein (zap) family acts as an antiviral protein against Sindbis virus and retro-viruses like Eco-luc (Gao *et al.*, 2002). Its antiviral activity is mediated through the disruption of viral messenger RNAs in the cytoplasm without affecting the levels of nuclear mRNA (Guo *et al.*, 2004). The human and

mouse zap proteins contain one CCCH-type zinc finger, one PARP catalytic domain and one WWE domain. The rat protein contains only the CCCH-type zinc finger and WWE domains. It has been shown that different CCCH zinc finger proteins interact with the 3' untranslated region of various mRNA (Gao *et al.*, 2002). A similar mechanism has been proposed for the zap interaction with viral RNA and subsequent exosome recruitment to degrade the mRNA (Gao *et al.*, 2002). The zap gene (known as ZC3HAV1 in human and mouse) of human is located at chromosome 7 and is flanked by the other genes that are also CCCH-type zinc finger proteins. The mouse gene maps to chromosome 6 and is also flanked by CCCH-type zinc finger protein genes. The rat gene is located on chromosome 4.

The promoter regions of zap genes were analyzed. The zap promoter region (**Supplementary Table 6.2**) contains a high number of motifs that are typically recognized by TFs of NHRs. Motifs 1, 2, 5, 6, 8, 9 and 14 correspond to binding sites of NHRs. Also, motifs that are associated with immune related TFs were identified. Motif 1 corresponds to E12, E47, c-Ets-2, LyF-1, USF-1. Motif 11 is associated with NF-1. Motif 19 corresponds to binding sites for NF-AT1, NF-1 and Ftz. Motif 20 corresponds to TFBSs for MTF-1. MTF-1 is known as metal-regulatory transcription factor 1. It is required for the basal transcription of metallothionein I and II genes (Heuchel *et al.*, 1994). It binds to metal response elements (MREs), which are related to Sp1. Heavy metals, but also oxidative stress (H_2O_2) and hypoxia can lead to increased MTF-1 activity and metallothionein expression (Zhang *et al.*, 2003) (Murphy *et al.*, 1999). MTF-1 has also been reported to regulate lymphocyte production (Wang Y. *et al.*, 2004).

6.4.4.1 Zap family promoter model

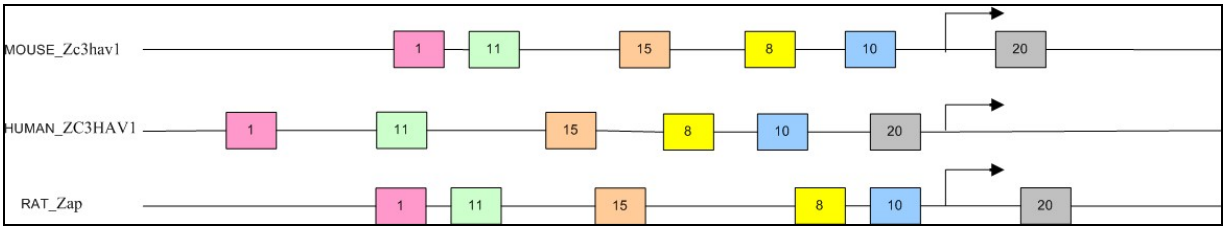
A motif arrangement for the zap promoter region that is common to all three species was identified (**Table 6.4, Figure 6.4**). The motif arrangement is in the order of 1-11-15-8-10-20 motifs. This arrangement corresponds to the following TFs, motif1 (Alfin1, RXR-alpha, VDR, E12, E47, MyoD, myogenin, EMF1, EMF2, EMF3, EMF4, Myf-5, c-Myc, USF2, CAN, E2A, DEP2, HEB, Ac, AS-C T3, Da, Sc, Sn, CLIM2, GATA-1, Lmo2, Tal-1, USF-1, NeuroD, NEUROD, LVa, PR B, AR, GR, c-Ets-2, ESE-1, HELIOS, LyF-1) - motif11 (NF-1, TGGCA-binding protein) - motif15 (Unknown) - motif8 (LyF-1, RXR-beta, VDR) - motif10 (Unknown) - motif20 (MTF-1). TFs in square brackets represent different TFs that potentially bind the associated motif.

The positional arrangement shows the presence of the NHR binding sites. The presence of NF-1 TFBS in both penk and zap families, suggests that transcripts of these families might be induced by steroid hormones that interact with NF-1 (Gronostajski, 2000). Zap protein is found in the liver and kidney at high levels. The presence of putative binding site for MTF-1, which is also localized in liver and kidney, suggests possible involvement of zap genes in metal regulation pathway or stress-related pathways like hypoxia.

Table 6.4: Motif arrangements in promoter region in mouse (F420004O17), human (HIX0007129.3) and rat (NM_173045) of zap family members
The species abbreviations are Hs: *Homo sapiens*; Mm: *Mus musculus*: Rn: *Rattus norvegicus*.

Species	Motif arrangement
Mm, Hs	5-7-9-16-4-13-2-1-11-15-8-19-10-17-20-6
Mm, Hs	1-11-15-8-10-20
Hs, Rn	1-11-15-8-10-20-18
Mm, Hs, Rn	1-11-15-8-10-20

Figure 6.4: Conserved Zap motif organization in mouse, human and rat



The numbers (i.e. 1) in the boxes refer to different motifs. The grey line represents the analyzed (-1000, +200) promoter sequence. This figure is a graphical representation of the common motifs across sequences determined by the DMB program. It is not drawn to scale.

6.5 Discussion

The author has shown that using *ab-intio* approach of finding significant motifs in promoters it is possible to generate models out of the motifs that could identify potentially co-regulated genes in a large data set. This approach does not require prior knowledge of the TFBSs that are present in the promoter regions. It is purely based on finding statistical over-representation of motifs across a set of sequences. The TFBS modules have been created from promoter regions of genes of orthologs, but the concept of phylogenetic footprinting has not been applied. The reason is that though phylogenetic footprinting may help in detecting functional binding sites, it restricts the region of motif search to only the conserved regions. In this study, motifs detected in the promoter regions were spread across the entire promoter region and not restricted to only the conserved regions between the orthologs. It was observed that with *ab-initio* search- known TFBSs motifs were also detected in non-conserved region of the promoters of AMP genes thus allaying the need for phylogenetic footprinting approach.

Some putative motifs in defa1, defa5, penk and zap promoter models are good candidates for experimental validation. For example, motif 12 of DEFA5 promoter model represents a potential binding site for LF-A1, GATA-1, COUP-TF2, NKX2-1, NF-E3. Motif 10 of DEFA1 represents putative binding sites for CCAAT-binding factor, NF-Y represented by motif 10. Motif 20 of DEFA1 represents Zic2. Many of the new motifs are located in the proximity to transcriptional start sites, suggesting that they have a role in mediating the binding of bona fide trans-factors.

6.6 Conclusion

The author was able to find a set of TFBS motifs that were present across all orthologous genes of the three AMP gene groups (alpha defensin, penk, zap). Promoter models on orthologs of three AMP gene groups were created. Both alpha defensin and penk had known TFBS and new TFBS motifs in their promoter models. Alpha defensin models contained known TFBS motifs like C/EBPalpha and some new motifs. Examples of new motifs are Zic2 (motif 20 of DEFA1), LF-A1, GATA-1, COUP-TF2, NKX2-1, NF-E3 (motif 12 of DEFA5). Penk promoter model had known TFBS motifs like motif 3 (GR), motif5 (AP-2) and new motifs such as motif 1 (RXR-alpha, LXR-alpha, ERRalpha1) and motif 13 (DSF, GCN4, COUP-TF1, RAR-beta, RXR-alpha, RAR-alpha1, TLX, Pax-2.1).

Zap promoter model indicates a linkage of zap, with NHRs and metal regulatory transcriptional control of innate immunity and oxidative stress. As stated in section 6.1 (Introduction), finding TFBS modules can lead us to predict other genes that are

responsive to the same set of TFs. Chapter 7 exemplifies this hypothesis using the alpha defensin promoter models.

Part III: Chapter 7: Implicated gene regulatory networks in AMPcg activities

Work spares us from three evils: boredom, vice, and need.
(Voltaire)

7.1 Introduction

In lower organisms, AMPs function merely as antibiotics by permeabilizing the cell membranes and lysing the invading microbes. However, during evolution these peptides have become multi-functional molecules acting in complex gene networks of higher organisms with additional properties like playing role as a mitogen or, taking part in adaptive immune responses (Kamysz *et al.*, 2003). Hence, it is likely that AMP genes are a part of more than one transcriptional co-regulation network. In support of this statement, experiments have revealed that TFBSs like USF2 and (Nicolas *et al.*, 2001), NKX2.2 (Wei *et al.*, 20051) are present in the promoter regions of some of the AMPcgs, are also involved in regulating the expression of non-immune gene pathways. The objective of this study has been to identify for the two different alpha defensin gene groups, candidate co-regulated genes that could be part of the same transcription regulation networks or otherwise be members of common activation phenomenon.

From the previous chapter (Chapter 6), two gene groups of alpha-defensins (alpha defensin 1 and alpha defensin 5) were selected that represent different cell origin. Alpha defensin 1 genes are expressed in the neutrophils whereas alpha defensin 5 genes are specific for paneth cells. For alpha defensin 5, human and mouse ortholog sequences were taken for the study. For alpha defensin 1, only human paralogs (defa1, defa3) were considered for promoter modeling. Defensins of neutrophil origin in human do not have corresponding mouse orthologs originating from neutrophils, thus alpha defensin 1 gene group was restricted to human sequences only. Using the shared TFBS organization

modules the human promoter dataset was scanned in search for genes that have a similar modular organization of elements in their promoters as that of the parent set of AMP genes. The predicted gene hits were then checked against co-expressed genes with the parent AMP genes that were extracted out of gene expression data.

7.2 Background

7.2.1 Gene regulatory networks (GRNs)

One gene can affect the expression of another gene by binding the gene product (protein such as TF) of one gene to the promoter region of another gene. Looking at more than two genes, regulatory networks can be referred as the regulatory interactions between the genes. Central to the computation of gene regulatory network (GRN) are DNA recognition sequences (TFBSs) with which TFs associate. When active transcription factors associate with the promoter region of target genes, they can function to specifically repress (down-regulate) or induce (up-regulate or activate) synthesis of the corresponding RNA. The immediate molecular output of a GRN is the constellation of RNAs and proteins encoded by network target genes. The resulting cellular outputs are changes in the structure, metabolic capacity, or behavior of the cell mediated by new expression of up-regulated proteins and elimination of down-regulated proteins. When creating a GRN, genes can be viewed as nodes in the network, with input being proteins such as TFs and outputs being the level of gene expression (Veiga *et al.*, 2006). The node itself can also be viewed as a function, which can be obtained by combining basic functions upon the inputs.

Mathematical models of GRNs have been developed to allow predictions of the models to be tested. Various modeling techniques have been used, including boolean

networks (Klamt *et al.*, 2006, Chaves *et al.*, 2005), Petri net (Matsuno *et al.*, 2000), Bayesian networks (Werhli *et al.*, 2006) and sets of differential equations (Chen *et al.*, 2005). Conversely, techniques have been proposed for generating models of GRNs that best explain a set of co-expressed genes (Dohr *et al.*, 2005).

The key yet unsolved problem with GRNs is identification of genes that form a particular network. This chapter approaches the generation of GRNs based on the later technique as discussed in the previous paragraph in the context of specific alpha defensins gene groups in human. The author shows a plausible approach to find genes that are part of the same GRN on the assumption that they a) share the same promoter model and b) are also co-expressed with the alpha-defensin 1 and 5 genes. TFs that bind motifs in the common promoter model are likely to be among the key drivers of co-expression of genes in the network.

7.2.2 Examples of known gene regulatory networks in AMPs (defensins)

Alpha defensins originating from the paneth cells exhibit numerous non-antimicrobial functions such as regulation of cell volume, chemotaxis, mitogenicity, and inhibition of natural killer cell activity (Ouellette, 1997). Mouse cryptidins 2 and 3 when administered apically can reversibly stimulate human T-84 intestinal epithelial cells to secrete chloride ion. This indicates that alpha defensins of paneth cells not only are components of the immune network of the crypt lumen but also influence the environment of the lumen by influencing other functional networks.(Ouellette, 1997).

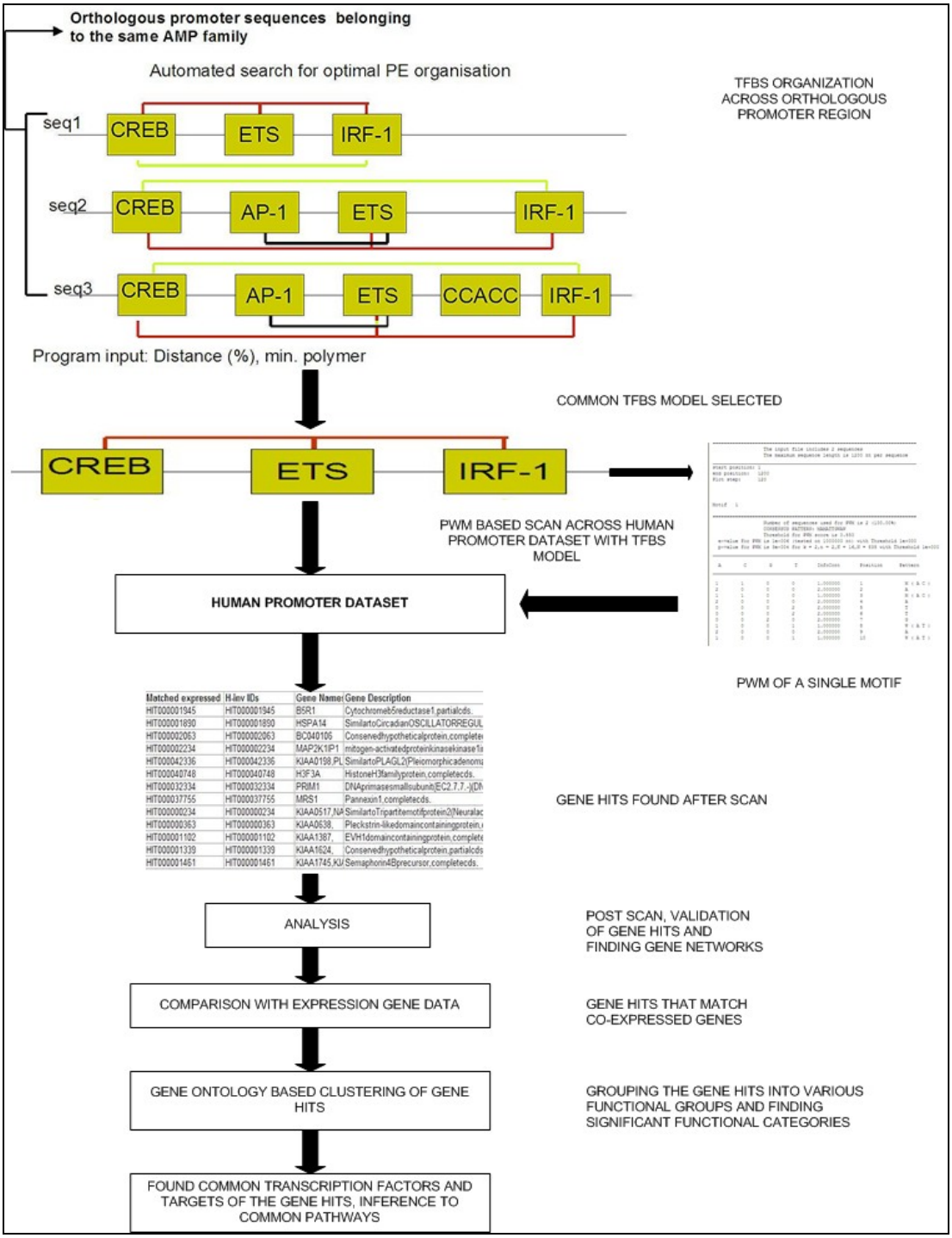
Toll receptor (TLR) -mediated activation of AMPs specifically defensins is the most well-studied pathway (Froy, 2005). Many of the TLRs are involved in activation of defensin sythesis such as TLR2, TLR3, TLR4, TLR5, TLR6 and TLR9

(Froy, 2005). TLR mediated activation of alpha defensins takes place in natural killer (NK cells) (Chalifour *et al.*, 2004). CD56+CD3– NK cells and some CD56+CD3+ T lymphocytes constitutively express alpha defensins (HNP-1, HNP-2 and HNP-3). NK cells CD56+CD3– and CD56+CD3+ are stimulated by the outer membrane protein A from *Klebsiella pneumoniae* and flagellin, that are the ligands of TLR2 and TLR5, respectively (Froy, 2005). This results in intracellular up regulation and secretion of alpha defensins. This phenomenon is different from the synthesis of alpha defensins in neutrophils which are constitutively produced and stored in phagosomes.

7.3 Materials and Methods

Figure 7.1 Workflow of generation of promoter models, scan across promoter dataset and analysis of gene hits

(The motifs shown in this diagram are examples, not representative TFBS models of the alpha defensin promoters)



7.3.1 Preparation of the promoter dataset

Preparation of the human target promoter data set to be scanned by the promoter models was done using cDNA data from H-invitational database, as well as Tag cluster groups from Fantom3 collection (ftp://fantom.gsc.riken.jp/FANTOM3/boundary_set/end5_clusters.txt.gz). cDNAs from H-invitational dataset were compared using the BLAT program against the human genome HG17 that was downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/chromosomes/). Only those cDNAs that satisfied 95% identity and 90% of the sequence length mapped to the human genome were chosen. The regions covering (-3000, +3000) inclusive of the gene were then extracted. Using information from the Tag cluster (TC) data, cDNAs were chosen based on their chromosome, strand, TSS location and number of tags. To chose the most accurate TSS position the following approach was implemented.

If the first 5' nucleotide of the CAGE tag or 5’ ditag (http://fantom31p.gsc.riken.jp/cage_analysis/export) coincided with the first 5′ nucleotide of the full-length cDNA (http://fantom.gsc.riken.go.jp/download.html), the TSS determined by this tag was selected. In cases when this condition did not hold, we selected TSSs where it we had information of a representative TSS location from a tag cluster that has at least ten tags, the representative TSS is supported by at least six tags, and there is at least one other piece of transcriptional evidence associated with this tag cluster (expressed sequence tag, full-length cDNA, or long SAGE; http://fantom.gsc.riken.go.jp/download.html). The resultant dataset had in total 10,255 sequences of human.

7.3.2 Quality assessment of the promoter models

In order to decide on the quality of the models, the sensitivity and specificity of the models were calculated. The models were validated using "leave-one-out-cross validation method (Nason, 1996). To determine sensitivity of the models the following procedure was applied. The DEFA1 training set had three promoter sequences, which consisted of DEFA1 (human), DEFA3 (human) and MNP1A (monkey). Two of the promoter sequences were used to generate the promoter model and the third sequence excluded. Then the promoter model was tested on the excluded sequence. This was done with each of the sequence in the training set. The sensitivity of each test was calculated. The average sensitivity of the model has been reported in (SupplementaryTable 7.1). Similar validation was done for DEFA5 model. The training set had 3 sequences that included the DEFA5 human, defcr3 mouse and defcr2 mouse (ortholog of DEFA5 human). The average sensitivity of DEFA5 model emerged as 100% (**Supplementary Table 7.1**). The cross validation was done with a small number of sequences in the training set as only orthologous genes having quality promoter sequence were chosen.

Specificity of the models was determined with a slightly different approach. All the three promoter sequences of DEFA1 training set were used to create a model. This model was applied to a test set of 18 AMP promoter sequences from different families. The model was able to pull out all the true positives ie. orthologs of DEFA1, DEFA3 in the test set and also some false positives.

Thus, the specificity of this model with three sequences was 45%. The results are shown in (**Supplementary Table 7.1)**. To make the model more specific, only two

sequences DEFA1, DEFA3 were taken into consideration. DEFA5 model pulled out only the DEFA5 related sequences from the test set. DEFA5 model showed 100% specificity.

7.3.3 Scanning the promoter data

When a suitable model could be generated that fulfilled the criterion as discussed in Chapter 6, (Section 6.2), it was chosen to scan the promoter data set for human. Position weight matrices (PWM) for each of the motifs (generated in the DMB results) contained in an AMP promoter model were used to scan the target human promoter data set of length 3000 upstream and 3000 downstream. The length of 3000 downstream was taken as many genes have multiple TSSs, which are spread across adjacently over a region (Bajic *et al.*, 2006). Not all TSSs in the region are functional and it is the functional TSS, which helps to determine the actual starting point of the promoter region. To take into account the entire TSS region, the length of 3000 downstream was considered.

The initial scanning of the promoter dataset was done based on a threshold which was the same as the threshold set to search motif using DMB (see section 2). If the threshold returned too few hits (less than 4 hits) then it was lowered to allow matching of the matrix model to ~250–300 genes, that would be a manageable number of genes to do further analysis. Only those promoters of the human target data set emerged as hits if the matched model contained all the motifs present in the same order as within the promoter model and were above or equal to the set matrix score threshold.

7.3.4 Comparison with expression data

To validate the model matches and assess if the predicted genes are likely to be part of the co-regulation network, hits were compared to co-expressed genes extracted from microarray expression data. The sources of microarray data were UCSC Expression (GNF Atlas 1) http://genome.ucsc.edu/cgi-bin/hgNear, NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/) and Standford microarray database (http://genome-www5.stanford.edu/). From UCSC genome browser and Stanford microarray database data of co-expressed genes was obtained for the parent AMP genes (DEFA1, DEFA5). From GEO, normal human tissue expression profile (HG-U95A), GDS422 consisting of 12 different tissue types were chosen and GDS260 representing data derived from pathogen exposure (*Leishmania major, Leishmania donovani, Toxoplasma gondii, Mycobacterium tuberculosis, Brugia malyi*) and immune response was taken. Using Pearson correlation coefficient, the genes that had similar gene expression profiles to the DEFA1, DEFA3, DEFA5 genes were selected. The cutoff of the correlation coefficient was above 60%. The selected genes comprised the co-expression data that was compared with promoter model predicted genes. The sample points with detection call = absent were not considered.

Collection of gene expression data from various experiments yielded a set of genes, which are co-expressed with the parent AMP genes. These set of genes were compared to the predicted list of genes produced by the scan and the common genes found were grouped as matched genes.

7.3.5 Statistical significance of predicted genes from the scan

All the predicted genes from a single model were matched to co-expressed genes as described in section 5 (section 7.3.4). The p-value was calculated for the enrichment in genes that are potentially co-regulated with each of the parent AMP gene (DEFA1, DEFA5). The p-value was determined using the hypergeometric distribution and the right-side Fisher's exact test (Fisher, 1922) and was corrected by the Bonferroni method **(Supplementary Table 7.2).**

7.3.6 Analysis based on Gene Ontology

To classify the promoter model hits according to function GO terms were extracted based on biological process and molecular function. FATIGO (http://www.fatigo.org/) (Al-Shahrour *et al.*, 2004) facilitated extraction and clustering of genes based on GO terms and pathways. The GO terms of the co-expressed genes were compared to GO terms of the model matches* to identify groups of genes, which have common GO categories. This is an indirect comparison approach that indicates that possibly the predicted gene set has genes which have same function as the genes from microarray data set and hence are probably co-regulated with the AMP genes of interest. The significance of this comparison of the GO terms in the two sets (predicted versus experimentally found co-expressed genes) has been statistically computed using Fisher's exact test and the p-values are corrected for multiple testing (Al-Shahrour *et al.*, 2004).

Figure 7.1 shows the schema for the promoter model scan and the post scan analysis

* model matches are the genes that emerged from promoter model scanning of the promoter dataset and has been use interchangeably with the word gene hits in this thesis

7.3.7 Finding common co-regulators and targets for the candidate network genes

The author identified common regulators and targets of gene hits through literature search. This was done to in order to decipher whether these genes are already known to be component of common pathways.

7.4 Results and Discussion

Using promoter models from different alpha-defensin groups and scanning 10,255 human promoters, the author identified sets of human genes that are likely to be co-regulated with parent alpha defensin groups. To validate the gene hits as plausible co-regulated gene candidates for a particular AMP gene, the following functional and regulatory based comparisons were done:

a. The gene hits were compared to co-expressed genes of AMPs to identify genes that are present in both datasets,
b. Functional comparison based on GO terms was undertaken between the co-expressed gene group and the predicted gene hits to observe similarities,
c. Gene hits that could not be identified in co-expressed gene data were grouped as unmatched gene hits which are possibly novel co-regulated genes,
d. The novel genes were compared and grouped alongwith the parent AMP gene based on the functions (GO based)
e. The novel genes were also compared with respect to function with the gene hits that emerged in the co-expressed gene data (matched gene hits) and,

f. Common regulators, transcription factors and downstream targets were looked at to deduce underlying common factors between the gene hits which potentially causes them to be co-regulated.

The following sections demonstrate the above points in details.

7.4.1 Alpha defensins

Alpha defensin (HNP1-4) also known as neutrophil defensins are synthesized constitutively by the bone marrow precursors of neutrophils (Selsted and Ouellette, 2005). The neutrophil defensins are then packaged in azurophil granules of neutrophils and comprise 30–50% of azurophil granule protein (Ganz *et al.*, 1985, Ganz, 1987, Rice *et al.*, 1987). The azurophil granules then fuse with phagocytes where they kill endocytosed microbes. Alpha defensins are thus only secreted when the neutrophils are stimulated. Alpha defensin 5 is an enteric defensin expressed mainly in the paneth cells and are constitutively produced (Cunliffe, 2003). Paneth cell alpha defensins are released when the cells are stimulated by cholinergic agonists and prokaryotic microbial antigens. HNP-1 (DEFA1 gene product) has been also been found to be expressed in by NK and T cells (Yang D. *et al.*, 2000). They are present in blood, bone marrow, plasma, spleen and thymus. Alpha defensin 1 is an antimicrobial peptide that is chemotactic for T cells and inhibits classical complement pathway (van den Berg *et al.*, 1998). It inhibits adenoviral infection and may play a role in tumor cell proliferation (Bastian and Schafer, 2001, Muller *et al.*, 2002 , Aarbiou *et al.*, 2002). Alpha defensin 5 is an antimicrobial and antifungal agent that is associated with, nasal polyps (Frye *et al.*, 2000), inflammatory bowel disease (Schmid *et al.*, 2004) and Crohn's disease (Wehkamp *et al.*, 2005). It is

found in colon, female reproductive tract, ileum, intestine, jejunum, small intestine, stomach and urogenital tract.

The specific regulatory elements and pathways that regulate alpha defensin synthesis and release in different tissues have not been well characterized. It was observed that promoter regions of pairs of defensin genes from the same site of expression, for example HNP1 and HNP4, and HD5 and HD6 (paneth cell specific) reveal marked similarities even in cases where the peptide sequence is highly divergent (Mallow *et al.*, 1996). Currently, existing knowledge about the promoter regions of human alpha defensins implies that they have binding sites for myeloid transcription factors that are essential for their transcription in HL-60 myeloid cell line (Ma *et al.*, 1998), but otherwise no detailed study of alpha defensin promoter structure has been published. This study provides the first more detailed analysis of these promoters. To date, no TFBS module or *cis*-regulatory module has been published for alpha defensins.

7.3.1 Scanned gene hits with alpha defensin models

The promoter model for alpha defensin 5 identified 240 unique promoters in the human promoter data set with a threshold of 0.77 (S**upplementary Table 7.3a, Supplementary Table 7.3b**). Out of 240, 177 gene hits matched experimentally found co-expressed data (73.75%). The co-expressed gene data set with DEFA5 consisted of 729 genes collected from various experiments (see section 7.3.4). However, out of 729, only 226 genes had their promoters in the human promoter data set. To determine the significance of match of the 177 gene hits with co-expressed data, statistical test was done (section 7.3.5). This yielded a Bonferroni corrected p-value of 1.765e-071 (**Supplementary Table 7.2**), which

indicated that predictions of genes that are co-expressed by DEFA5 based on promoter model is very good.

For alpha defensin 1 scanning the human promoter dataset with a threshold setting of 0.65, yielded 104 hits (**Supplementary Table 7.4a, Supplementary Table 7.4b**). The collection of co-expressed genes for alpha defensin 1 and alpha defensin 3 was 472. Promoters for 51 genes were found in the human promoter dataset. Out of 51 promoters, 17 hits emerged in the prediction list. Hence, 17 genes coincided with experimentally found co-expressed genes with DEFA1 and DEFA3 genes. Similar statiscal test was carried out in this case. The p-value of these predicted genes (17 genes) was 3.72E-18.

It was observed that the number of genes that emerged as hits using the promoter model for DEFA1-3 was less, although a lower threshold was used for scanning as compared to DEFA5 genes. DEFA1 and DEFA3 promoter regions are highly similar. Hence, the promoter model generated consisted of most of the 20 motifs in the same order in both the promoter regions of DEFA1 and DEFA3. However, having several motifs in a model makes the model highly restrictive. Hence, a optimal number (3–4) motifs should be used to have a promoter model. For DEFA1, DEFA3, there were several combinations that could be used. The author chose the model that had more number of motifs that corresponded to known TFBSs found in the promoter region of DEFA1-3. It is likely that although two of the motifs are known to occur in alpha defensin promoters, the combination of all the four motifs in the promoter model may have not been found in many promoters of the human data set and in DEFA1-3 co-expressing genes.

7.3.2 Some interesting gene hits from DEFA1and DEFA5 promoter model scan

Alpha defensins are known to be involved in adaptive immune pathways other than their main role as antimicrobial peptides of innate immunity. Some of the scanned gene hits for DEFA1, DEFA3 and DEFA5 have direct or indirect association with immune pathways. These gene hits have been discussed in the following paragraphs.

CX3CL1 was one of the gene hits that emerged from both DEFA1, DEFA5 promoter model scan. It is a membrane-expressed protein promoting cell-cell adhesion, which also is a soluble molecule inducing chemotaxis. It is known that besides having chemotactic property, some chemokines like CXCL4, CXCL9, CXCL10, CXC11, CTAP3, RANTES also have antimicrobial activity (Krijgsveld *et al.*, 2000 Durr and Peschel, 2002). CX3CL1 did not appear in the co-expressed data for either DEFA1 or DEFA5. However, its functional property of being associated with the immune pathway, and having common promoter elements makes it a probable co-expressed gene with DEFA1-3 and DEFA5.

DEFA1 gene hit, FKBP12 is an immunophilin, which plays a role in immuno-regulation and basic cellular processes involving protein folding and trafficking. It complexes with immunosuppressor protein FK506 and inhibits calcineurin which is involved in activation of NF-kappaB (Odom *et al.*, 1997). Inhibition of calcineurin in fungal pathogen *Cryptococcus neoformans* adversely effects virulence (Odom *et al.*, 19975). It also interacts with several other intracellular signal transduction proteins including type I TGF-beta receptor. Model match for DEFA1, VSIG2 is a member of the immunoglobulin domain cell adhesion molecule (cam). Another gene hit of DEFA1 scan, PSMB8 is involved in the process of antigen presentation (Schwarz *et al.*, 2000). Model hits CPNE6 and CPNE4 belong to the C2 domain family which are Ca2+-dependent membrane-

targeting module found in many cellular proteins involved in signal transduction or membrane trafficking. C2 domains are unique among membrane targeting domains in that they show a wide range of lipid selectivity for the major components of cell membranes, including phosphatidylserine and phosphatidylcholine. CCND2, which is a G1/S phase specific cyclin was one of the hits that coincided with expression data for DEFA1, DEFA3 genes. CCND2 and alpha defensins are over-expressed in colon cancer when induced by a carcinogen, PhIP (Fujiwara *et al.*, 2004). The model matches with DEFA1 promoter model are reported in (Supplementary Table 7.4a) and (Supplementary Table 7.4b). INS (insulin) emerged as another gene hit that is implicated to be expressed in thymus and induces tolerance in $CD8^+$ T cells (Ma *et al.*, 2000). Insulin is involved in alpha-beta T-cell activation (Ma *et al.*, 2000).

DEFA5 scan result yielded seven gene hits were grouped under immune response namely HLA-DMA, ILF2, G10P1, TTF, IFI30, AIF1, DHLAG. HLA-DMA, AIF1, DHLAG. TTF, DHLAG are involved in lymphocyte differentiation (Table 7.1). CKLFSF6 is another gene hit that belongs to the chemokine-like factor gene superfamily, which is a novel gene family and has properties that indicate that it has chemokine and chemotaxis activity (Han *et al.*, 2003).

Table 7.1 is a summary of DEFA1 and DEFA5 gene hits that have been discussed in the previous paragraphs. **Figure 2a** and **Figure 2b** show the regulatory networks for the genes listed in **Table 7.1**. The networks were created using Ingenuity system software (www.ingenuity.com).

In the DEFA1 regulatory network (**Figure 2a**), TNF and IL-15 are the key regulators. TNF controls expression of INS and is also regulated by INS (Hostens *et al.*,

1999, Iida *et al.*, 2001). Other genes such as CCND2, CX3CL1, PSMB8 are regulated by TNF (Banno *et al.*, 2004, Li *et al.*, 2002), Banno *et al.*, 2004). IL-15 regulates PSMB8, DEFA1 (Tourkova *et al.*, 2005, Liu *et al.*, 2002). DEFA1 is known to bind to SERPING1 and inhibit the classical complement pathway (van den Berg *et al.*, 1998). SERPING1 decreases the mRNA expression of TNF-alpha in mouse (Liu *et al.*, 2003). Whether the DEFA1, SERPING1 interaction has any direct effect on TNF-alpha expression is unknown.

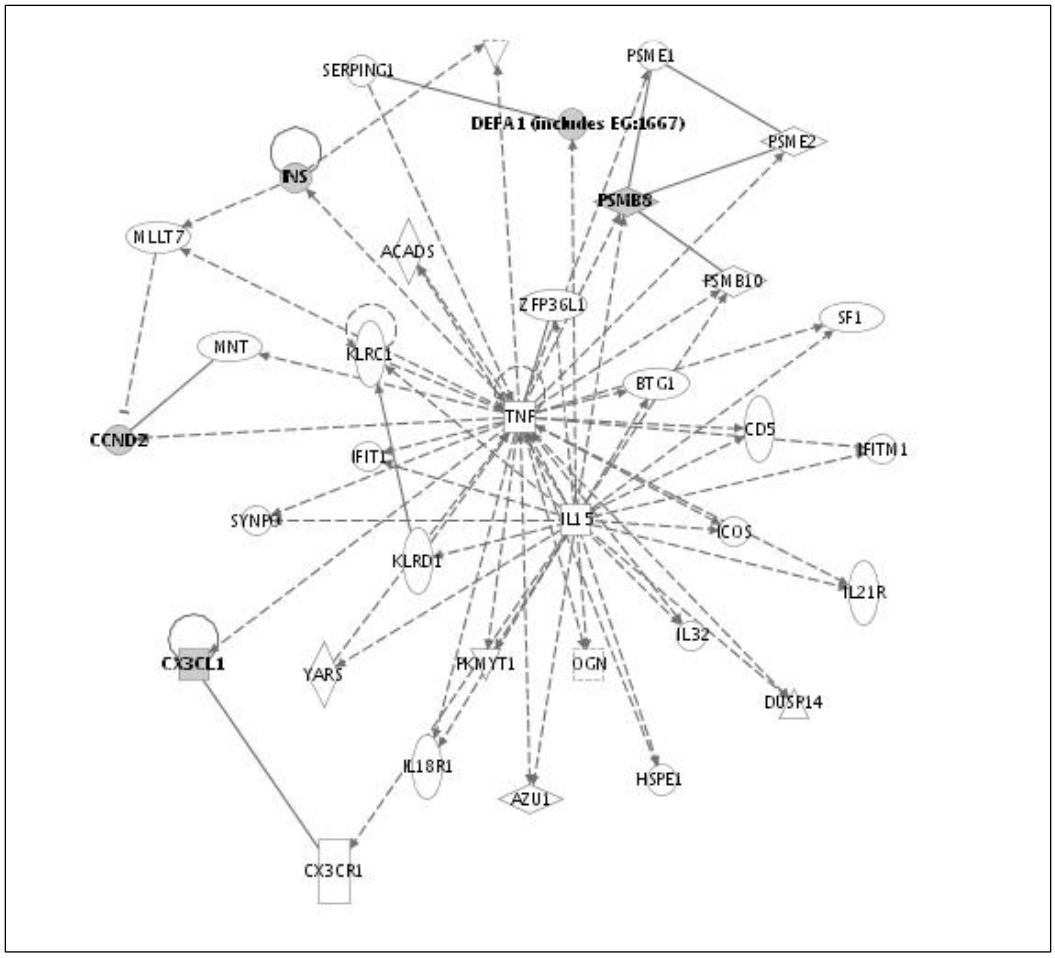
For the DEFA5 network (**Figure 7.2b**), IFNG (interferron, gamma), PTEN (phophatase and tensin homolog) and ILF3 (inteleukin enhancer binding factor, 3) are at the core of the network. IFNG indirectly regulates many of the genes that have been listed in **Table 7.1** for DEFA5, such as CX3CL1 (Ludwig *et al.*, 2002), AIF1 (Autieri *et al.*, 2000), IFIT1 (Okumura *et al.*, 2003), HLA-DMA (Muczynski *et al.*, 1998), CD74 (DHLAG) (Cao *et al.*, 2000). It appears that PTEN indirectly acts as a negative regulator for DEFA5 expression. It decreases binding of DNA and a protein-protein complex consisting of human beta catenin (ctnnb1) and of human Tcf (Persad *et al.*, 2001) that activate DEFA5 expression (Schwartz *et al.*, 2003). PTEN negatively regulates IFI30 (Matsushima-Nishiu *et al.*, 2001).

Table 7.1 Selected gene hits of DEFA1 and DEFA5

H-inv ID: Ids from the H-Invitational database

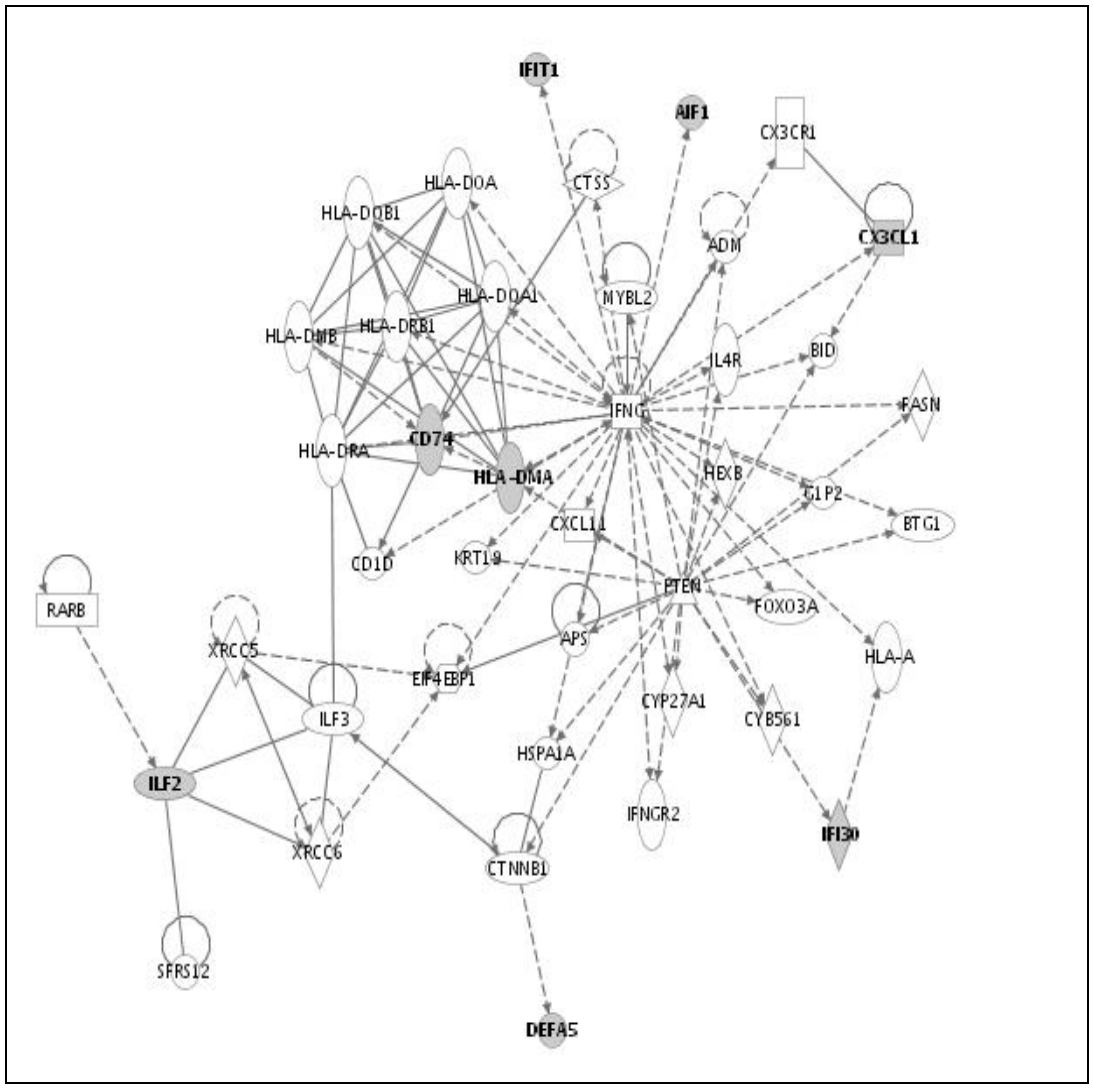
H-inv ID	Gene symbol	Gene Description	Pathway	Tissue origin
HIT000037490	CX3CL1	chemokine (C-X3-C motif) ligand 1; small inducible	Cytokine-cytokine receptor interaction	Brain,neuroblastoma
HIT000032247	FKBP12, FKBP1A	FK506-binding protein 1A (EC 5.2.1.8)	mTOR signalling pathway	Placenta,choriocarcinoma
HIT000033138	VSIG2	Immunoglobulin subtype domain containing protein, complete cds.		Colon,adenocarcinoma
HIT000030145	Y2,PSMB8	Similar to Proteasome subunit beta type 8 precursor (EC 3.4.25.1)		Skin,melanoticmelanoma.
HIT000038235	CPNE6	Copine VI (Neuronal-copine) (N-copine), partial cds.		Brain,hypothalamus
HIT000036673	CPNE4	Copine IV, complete cds		Brain,neuroblastoma
HIT000035146	CCND2	G1/S-specific cyclin D2, partial cds.	Cell cycle, Wnt signaling pathway, Focal adhesion, Jak-Stat signaling	Bonemarrow,chronicmyelogenousleukemia
HIT000032325	INS	Insulin precursor, complete cds.	Regulation of actin cytoskeleton, Insulin signaling pathway, Dentatorubropallidoluysian atrophy (DRPLA)	Pancreas
DEFA5				
HIT000035253	HLA-DMA	majorhistocompatibilitycomplex,classII,DMalphaprecursor;	Cell adhesion molecules (CAMs)	Skeletal Muscle
HIT000029571	ILF2	NF45protein,completecds.		Lung, small cell carcinoma
HIT000031192	CKLFSF6	chemokine-likefactorsuperfamily6(Homosapiens), completecds.		Ovary, adenocarcinoma
HIT000040025	G10P1	SimilartoInterferon-inducedproteinwithtetratricopeptide		Pancreas, Spleen, adult pooled
HIT000036609	TTF	Rho-relatedGTP-bindingproteinRhoH(GTP-bindingproteinTTF),		Primary B-Cells from Tonsils
HIT000038424	DHLAG	HLAclassIIhistocompatibilityantigen,gammachain(HLA-DR		Primary B-Cells from Tonsils
HIT000034389	AIF1	SimilartoAllograftinflammatoryfactor-1(AIF-1)(Daintain),		Prostate
HIT000039106	IFI30	interferon,gamma-inducibleprotein30preproprotein;		Skin, melanotic melanoma, high MDR.
HIT000037490	CX3CL1	chemokine(C-X3-Cmotif)ligand1;smallinduciblecytokine	Cytokine-cytokine receptor interaction	Brain, neuroblastoma

Figure 7.2a Network of DEFA1 and genes that resulted from the promoter model matching



The dotted lines indicate indirect relationships and the other lines indicate direct relationship. The grey shaded circles and boxes are the genes of interest found by the promoter model scan. Rhombuses: enzymes, rectangle: ligand dependent nuclear receptor, oval horizontal circles: transcription regulator, oval vertical circles: transmembrane receptor.

Figure 7.2b: Network of DEFA5 and genes that resulted from the promoter model matching



The dotted lines indicate indirect relationships and the other lines indicate direct relationship. The grey shaded circles and boxes are the genes of interest found by the promoter model scan. Rhombuses: enzymes, rectangle: ligand dependent nuclear receptor, oval horizontal circles: transcription regulator, oval vertical circles: transmembrane receptor

7.4.4 Comparison of gene hits with co-expressed genes for DEFA1, DEFA3

The author compared the predicted gene hits with co-expressed gene data for human DEFA1 and DEFA3 and found eleven genes (CCND2, ABHD2, TMED9, ARB2, FKBP12, MARS, MEA1, CNOT2, PIAS2, CASP5, RSU) that overlapped with the co-expressed gene data sets of DEFA1 and DEFA3. Another set of model matches (UGT2B11, DDX23, MARCH5, ZNF33A, VSIG2, PSMB8) had similar protein domains and GO functions to the co-expressed genes of DEFA1 and DEFA3 (UGT2B15, DDX27, MARCH3, ZNF167, VSIG4, PSMB4) respectively (**Supplementary Table 7.4a**). The gene hits that did not overlap with the co-expressed data set were grouped as "unmatched gene hits". These unmatched gene hits that had GO terms were compared with DEFA1, DEFA3 GO terms of co-expressed gene group to determine if they could be categorized under similar GO categories. Several GO categories were common and significant in both the unmatched gene group and the co-expressed gene group (**Supplementary Figure 7.1**). The common GO categories that came up were regulation of cellular physiological process, defense response, regulation of metabolism, signal transduction, primary, cellular and macromolecule metabolism, biosynthesis.

Next, all the gene hits from DEFA1, DEFA3 model with GO terms were compared with GO terms for entire co-expressed gene dataset for DEFA1, DEFA3. The results showed that biological processes like regulation of cellular and physiological process, cell communication, and response to stimulus had significantly comparable percentage of genes represented in both data sets. The significance of this comparison

was determined by p-values as shown in Supplementary **Figure 7.2**. The null hypothesis in this comparison was that the genes from the two groups (predicted versus co-expressed) fall under similar GO categories. This hypothesis was supported by the corrected p-values as shown in column 4 of **Supplementary Figure 7.2**. The high p-values indicated that the gene groups were not significantly different, therefore, the null hypothesis was true. Therefore, many of the predicted genes for DEFA1, DEFA3 appear to have similar functions to the experimentally derived co-expressed genes for DEAF1, DEFA3.

Comparison was also performed at molecular function level. The results indicated that, protein binding, nucleic acid binding, ion binding, oxido-reductase activity, ion transporter activity, hydrolase activity and transferase activity had significant number of genes represented from both data sets for these categories (**Supplementary Figure 7.3**).

7.4.5 Gene ontology based clustering for DEFA1 gene hits

For alpha defensin 1 group, gene hits were clustered based on GO biological process and molecular function. Clustering based on GO biological process showed 13 hits categorized under cellular metabolism and primary metabolism. MIP, FTL, ATP5S, CPNE6, SRPR, SEC5L1 are involved in the process of transport. Six genes (TAF11, MYST2, CCND2, CHD2, CNOT2, CCNI) came under regulation of cellular physiological process. APBB1IP, CX3CL1, FMOD, INS are involved in signal transduction processes. Three genes UGT2B11, MIP, CX3CL1 were categorized under the GO category response to external stimulus (**Supplementary Table 7.5a**). Analysis of all gene hits at GO molecular function level showed nucleic acid binding function as the

most significant function covering 10 hits. Next was function category of hydrolase activity with CHD1L, CHD2, DDX23, and SRPR under its category (**Supplementary Table 7.5b**).

7.4.6 Alpha defensin 5 gene hits

Alpha defensin 5 promoter model yielded 240 unique gene hits. The predicted gene list was divided into two groups as previously done with DEFA1 gene hits One group consisted of those genes that matched expression data (co-expressed gene data for DEFA5 human) and the other section had those that did not match (Supplementary Tables 7.3a, and 7.3b). GO based analysis of matched gene group with respect to biological process showed that the largest number of gene hits came under cellular, primary and macromolecular metabolism (**Supplementary Table 7.6a**). The next two categories that had many genes clustered under them were regulation of cellular physiological process, localization and transport. Seven hits, HLA-DMA, ILF2, G10P1, TTF, IFI30, AIF1, DHLAG were grouped under immune response.

GO based clustering in the unmatched group also had cellular, primary and macromolecular metabolism categories having the highest number of genes (**Supplementary Table 7.6b**).

GO biological functions like immune response, signal transduction, transport, cellular, primary and macromolecular metabolism, cell cycle, localization and a few other categories showed comparable number of genes between the matched group and unmatched group of predicted genes (**Supplementary Figure 7.4**). The significance of the comparison is depicted by the corrected p-values in (**Supplementary Figure 7.4**).

This perhaps indicates that although the genes in the unmatched category do not coincide with experimentally found co-expressed genes for DEFA5, they have similar functions to the co-expressed genes for DEFA5. Another observation was that many of the genes in both the matched and unmatched groups for alpha defensins came under non-immune categories like metabolism, transport and localization. These findings support the multifunctionality of defensins and their involvement in pathways other then innate immunity. In fact, some of these functions may be attributed to the tissue cell ontogeny and evolutionary adaptations. **Table 7.2** lists the significant GO categories for the unmatched (novel) gene hits of DEFA1 and DEFA5.

Table 7.2: The GO terms having the maximum number of novel (predicted gene hits not in the co-expressed gene data) gene hits from DEFA1 and DEFA5

DEFA1		
GO biological function	**DEFA1_umatched gene hits**	**No. of genes**
primary metabolism	FARS1 DPM1 CHD1L MTMR5 HRMT1L1 PECI KIAA0065 MYST2 H1F0 ALDR1 TMPRSS1 LSM6 GUK1 FKBP1A KIAA0929 KIAA0935 CPNE6 CHD2 MIZ1 TAF11 PDCD9 KBL NCL PFD4 RODH KIAA0060 ARAF INS	28
cellular metabolism	FARS1 DPM1 CHD1L MTMR5 HRMT1L1 PECI KIAA0065 MYST2 H1F0 NDUFS5 TMPRSS1 LSM6 GUK1 FKBP1A KIAA0929 KIAA0935 CHD2 MIZ1 TAF11 PDCD9 KBL NCL PFD4 RODH KIAA0060 ARAF INS	27
macromolecule metabolism	FARS1 DPM1 CHD1L MTMR5 HRMT1L1 MYST2 H1F0 ALDR1 TMPRSS1 LSM6 FKBP1A KIAA0935 CHD2 PDCD9 NCL PFD4 KIAA0060 ARAF INS	19
establishment of localization	NTT73 ATP5S BGP1 CX3CL1 FTL NDUFS5 CPNE6 SEC5L1 INS	9
signal transduction	BGP1 HRMT1L1 CX3CL1 FKBP1A KIAA0929 FMOD APBB1IP ARAF INS	9
regulation of metabolism	KIAA0065 MYST2 KIAA0929 CHD2 MIZ1 TAF11 INS	7
biosynthesis	FARS1 DPM1 GUK1 PDCD9 KBL RODH INS	7
transport	NTT73 ATP5S FTL NDUFS5 CPNE6 SEC5L1 INS	7
regulation of cellular physiological process	KIAA0065 MYST2 KIAA0929 CHD2 MIZ1 TAF11 CCNI	7
cell organization and biogenesis	MYST2 H1F0 PEX11G CHD2	4
nitrogen compound metabolism	FARS1 KBL KIAA0060 INS	4
catabolism	RODH KIAA0060 INS	3
cell cycle	MIZ1 CCNI DCTN3	3
regulation of organismal physiological process	CX3CL1 INS	2
cell-cell adhesion	BGP1 CX3CL1	2
positive regulation of cellular process	CX3CL1 FKBP1A	2
immune response	CX3CL1 INS	2
defense response	CX3CL1 INS	2
cell-cell signaling	CPNE6 INS	2
regulation of signal transduction	FKBP1A	1
DEFA5		
GO biological function	**DEFA5_unmatched_genes**	**No. of genes**
cellular metabolism	PCCX1 NPD002 VPS11 TRIM9 EFCBP1 ELE1 RPS27L RBM3 PSMA7 MXD3 H2AFX EPM2A USP39	13
primary metabolism	PCCX1 VPS11 TRIM9 ELE1 RPS27L RBM3 PSMA7 MXD3 H2AFX EPM2A USP39	11
macromolecule metabolism	VPS11 TRIM9 RPS27L RBM3 PSMA7 H2AFX EPM2A USP39	8
establishment of localization	NPD002 VPS11 CX3CL1 VIM TFIP11 KCNMA1	6

transport	NPD002 VPS11 VIM TFIP11 KCNMA1	5
regulation of cellular physiological process	PCCX1 YWHAG ELE1 MXD3 EPM2A	5
signal transduction	YWHAG TENC1 CX3CL1 ELE1 PDZK2	5
Table 7.2 continued		
regulation of metabolism	PCCX1 ELE1 MXD3 EPM2A	4
biosynthesis	EFCBP1 RPS27L EPM2A	3
positive regulation of cellular process	CX3CL1 ELE1	2
catabolism	PSMA7 USP39	2
cell organization and biogenesis	YWHAG H2AFX	2
cell-cell signaling	YWHAG KCNMA1	2
cell cycle	YWHAG H2AFX	2

7.4.7 Common regulators and targets of the predicted gene hits

Common transcriptional regulators and common targets for the gene hits for each of the gene groups (DEFA1, DEFA5) were looked at to understand the commonalities in transcription regulation of the predicted gene hits. These links between the gene hits and the common regulators and targets is substantiated by literature evidence.

For DEFA1 gene hits, ADP is involved in regulation of DEFA1, H1F0 (h1 histone family, member 0) and INS (insulin) (Paone *et al.*, 2002, Adamietz *et al.*, 1978, Petit *et al.*, 1989).

Model matches of DEFA1, INS (insulin), MARS (methionine-tRNA synthetase), H1F0 (h1 histone family, member 0), AKR1B1 (aldose reductase) are involved in cell differentiation function (**Table 7.3**), a function that is common with DEFA1, DEFA3. HNP1-3 (DEFA1, DEFA3) is known to be involved in mucin cell differentiation (Aarbiou *et al.*, 2004). The neutrophil defensins have mitogenic properties as demonstrated on epithelial cells and fibroblast (Murphy *et al.*, 1993). DEFA1 gene hits, INS, CCND2, CEACAM1, FKBP1A, TNPO1, NCL, SBF1, H1F0, CCNI also appear to be involved in various mitogenic functions (**Table 7.3**). Moreover, INS is known to

synergistically act with the defensins in the mitogenic process (Murphy *et al.*, 1993. Many of the DEFA1 gene hits such as NCL, MIP, INS, CEACAM1, PDCD4, CCND2, ARAF1, AKR1B1, H1F0 are regulated by protein kinase C. Interestingly, DEFA1 inhibits protein kinase C (PKC) activity in CD4+T cells (Chang *et al.*, 2005) causing inhibition of HIV-1 replication.

Model hits INS, H1F0 are reglated by various hormones like glucocoticoid, progesterone and thyroid stimulating hormone.

For the alpha defensin 5gene hits, GFAP, BGN, TXNRD1, CLG4A, NAALAD1, MMP2, FOLH1 and F3 have TNF (tumor necrosis factor) involved in their regulation. TNF-alpha expression is up-regulated by alpha defensins (Chaly *et al.*, 2000). A common regulator of some of the DEFA5 gene hits is interleukin 1-beta (IL1B). It is involved in regulation of FABP1, MMP2, F3 and AIF1. IL1B has been observed to stimulate MMP2 in cultured rat astrocytes. IL1B positively regulates F3, AIF1, IFI30 and negatively effects FABP1. DEFA5 gene hits CX3CL1, YWHAG, VIM and PSMA7 come under cytokine regulation.

Some of the DEFA5 gene hits have known regulatory effects on other gene hits. For example, the gene Claudin-2 is also involved in the formation of intestinal epithelial barrier and its gene expression is up-regulated when stimulated with interleukins (Kinugasa *et al.*, 2000). Claudin-2 (CLDN2) and Discoidin domain receptor 2 (DDR2) indirectly increase the protein activity of another DEFA5 gene hit, MMP2. Likewise, HSPA14 increases the protein secretion of CCL4. MLL binds to the promoter of HOXA9 and increases its mRNA abundance.

Table 7.3 gives a detailed overview of regulators and targets.

Transcription factors (TFs) that regulate many of these genes were also looked at. **Supplementary Table 7.7** gives a list of the common TFs found across the predicted gene hits. These TFs were found across various genes by implementing FATIGO *plus* analysis module for finding transcription factors (Al-Shahrour *et al.*, 2006). Some of the TFs found across several of these genes are immune system regulatory factors. HNF-1, CDX, Nkx2-5, GATA-4, LXR, PXR, CAR, COUP, RAR, Oct-1, and NF-kappaB are the commonly found TFs across both the matched and unmatched gene sets for DEFA5. NF-kappaB is a key regulatory transcription factor for genes involved in response to infection, inflammation, stress (Baeuerle and Henkel, 1994) , (Sica *et al.*, 1997) , (Quinlan *et al.*, 1999) , (Hiroi and Ohmori, 2003). GATA-4 is also known to be involved in regulation of immune system (Su *et al.*, 2004). Nkx2-5, GATA-4, HNF-1, CDX are the common TFs found in both the matched and unmatched predicted gene hits for DEFA1. HNF-3alpha, Evi-1 were the two other TFs that were found in genes that matched experimental co-expressed data set for DEFA1. Both Defa1 and DEFA5 gene hits have common TFs such as NF-kappB. Nkx2-5,GATA-4,CDX etc..

Table 7.3 Common regulators and common targets of DEFA1 and DEFA5 predicted genes

(*PMID: *Pubmed unique identifier*). ---> indicates regulation, ---+> positive regulation, -----| negative regulation

Defa1			
Type	**Nodes**	**Effect**	**References (PMID)***
Regulators			
ProtModification	ADP ---> DEFA1		12060767
ProtModification	ADP ---> H1F0		729572
MolTransport	ADP ---> INS		2686791
Regulation	glucocorticoid --+> INS	positive	11121405
Regulation	thyroid stimulating hormone --+> INS	positive	11473059
Regulation	progesterone --+> H1F0	positive	8187766
Regulation	PKC□INS	negative	11246878
ProtModification	PKC->H1F0		3028404
Regulation	PKC->MIP		2541249
ProtModification	PKC ---> NCL		10811822
ProtModification	PKC ---> CEACAM1		10754323
Regulation	PKC --+> PDCD4	positive	2752524
Expression	PKC --+> CCND2	positive	11120786
Regulation	PKC --+> ARAF1	positive	8621729
Regulation	PKC --+> AKR1B1	positive	12527382
Targets			
Regulation	INS	differentiation	11872678, 10385414
Regulation	DEFA1	differentiation	9352884
Regulation	MARS	differentiation	4331137
Regulation	H1F0	differentiation	1988682
Regulation	AKR1B1	differentiation	151810922
Regulation	DEFA3	differentiation	12871849
Regulation	INS	mitogenesis	10706096, 12183434
Regulation	CCND2	mitogenesis	11691826
Regulation	CEACAM1	mitogenesis	11694516
Regulation	FKBP1A	mitogenesis	11226255
Regulation	TNPO1	mitogenesis	9388191
Regulation	NCL	mitogenesis	12506112, 10811822
Regulation	SBF1	mitogenesis	12704202
Regulation	H1F0	mitogenesis	15694489
Regulation	CCNI	mitogenesis	11054536
Defa5			
Expression	TNF ---> GFAP		8622125
Regulation	TNF ---> BGN		11322893
Regulation	TNF --+> TXNRD1	Positive	14584040

Regulation	TNF --+>CLG4A	Positive	10233890
Regulation	TNF --+> NAALAD1	Positive	12744776
MolSynthesis	TNF --+> F3	Positive	9002957
Regulation	IL1B ---| FABP1	Negative	10477831
Regulation	IL1B ---> MMP2		8945720
Regulation	IL1B --+> F3	Positive	12429585
Regulation	IL1B --+> AIF1	Positive	10894811
Regulation	IFNG --+> IFI30	Positive	12215441

7.4.8 Commonality and differences between DEFA1 and DEFA5 gene hits

DEFA1 and DEFA5 genes belong to the same AMP family and hence it is expected that they would have similar functions and perhaps be involved in similar gene networks. Interestingly, the gene hits that emerged from both the models did not have a significant overlap. This instigates the curiosity to compare the gene hits of DEFA1 and DEFA5 and observe the commonalities and differences between them.

DEFA1 and DEFA5 gene hits (104, 240 respectively) were compared based on their GO terms and pathways. It was observed that metabolic activity, signal transduction, localization, biosynthesis, transport, cell death, neuro-physiological process, cell activation and immune response were the common GO function categories between DEFA1 and DEFA5 gene hits (**Figure 7.3, Supplementary Table 7.8**). However, some GO functions were exclusive to either group of gene hits as discussed later in this section. DEFA5 gene hits had involvement in more varied functions than DEFA1 gene hits. This could be due to the unequal number of gene hits emerging from the different model scans. DEFA5 gene hits appeared to be more involved in cell differentiation, organ development and cell growth functions compared to DEFA1 gene hits (**Figure 7.4, Figure 7.5, Supplementary Table 7.8**).

In terms of pathway level comparison, both DEFA1 and DEFA5 gene hits were involved in metabolic pathways such as prostaglandin and leukotriene, fructose and mannose, starch and sucrose, tyrosine, purine, galactose metabolism etc. Other pathways that had gene hits from both groups comprised of the WNT signaling pathway, MAPK signaling, insulin signaling and cell cycle. The GO comparison shows that genes from both alpha defensin groups are involved in signal transduction. Since GO function terms and pathways of a gene are interdependent, it is not surprising to see these signaling pathways emerge in the analysis of AMP co-regulated genes. MAPK signaling pathway is involved in innate immune responses as it is involved in activation of macrophages (Schorey and Cooper, 2003). WNT-signaling pathway has a important role to play in organ development, and dysregulated WNT signaling causes tumors. Recently its role has been implicated at several stages of lymphocyte development and in the self-renewal of haematopoietic stem cells (Staal and Clevers, 2005). The GO analysis of DEFA1 and DEFA5 gene hits show that many of them are involved in organ development (**Figure 7.4, Figure 7.5, Supplementary Table 7.8**). Insulin signaling pathway is indirectly involved in the regulation of the immune system (McKenzie *et al.*, 2006). Pathways that affect innate immunity have not been studied as well as that of adaptive immunity. An analysis of this kind indicates that many known pathways that in the current knowledge have no link to immune related mechanisms may perhaps be involved in direct or indirect ways in regulation of the latter.

Within the data analyzed, some pathways appeared to be exclusive to a single gene group such as the Jak-Stat signaling pathway that involved gene hits of the DEFA1

group only. DEFA5 gene hits were involved in calcium signaling pathway. **Table 7.4** shows the comparison of DEFA5 and DEFA1 gene hits with respect to pathways.

Figure 7.3: GO biological functions that are common between DEFA1 and DEFA5 gene hits

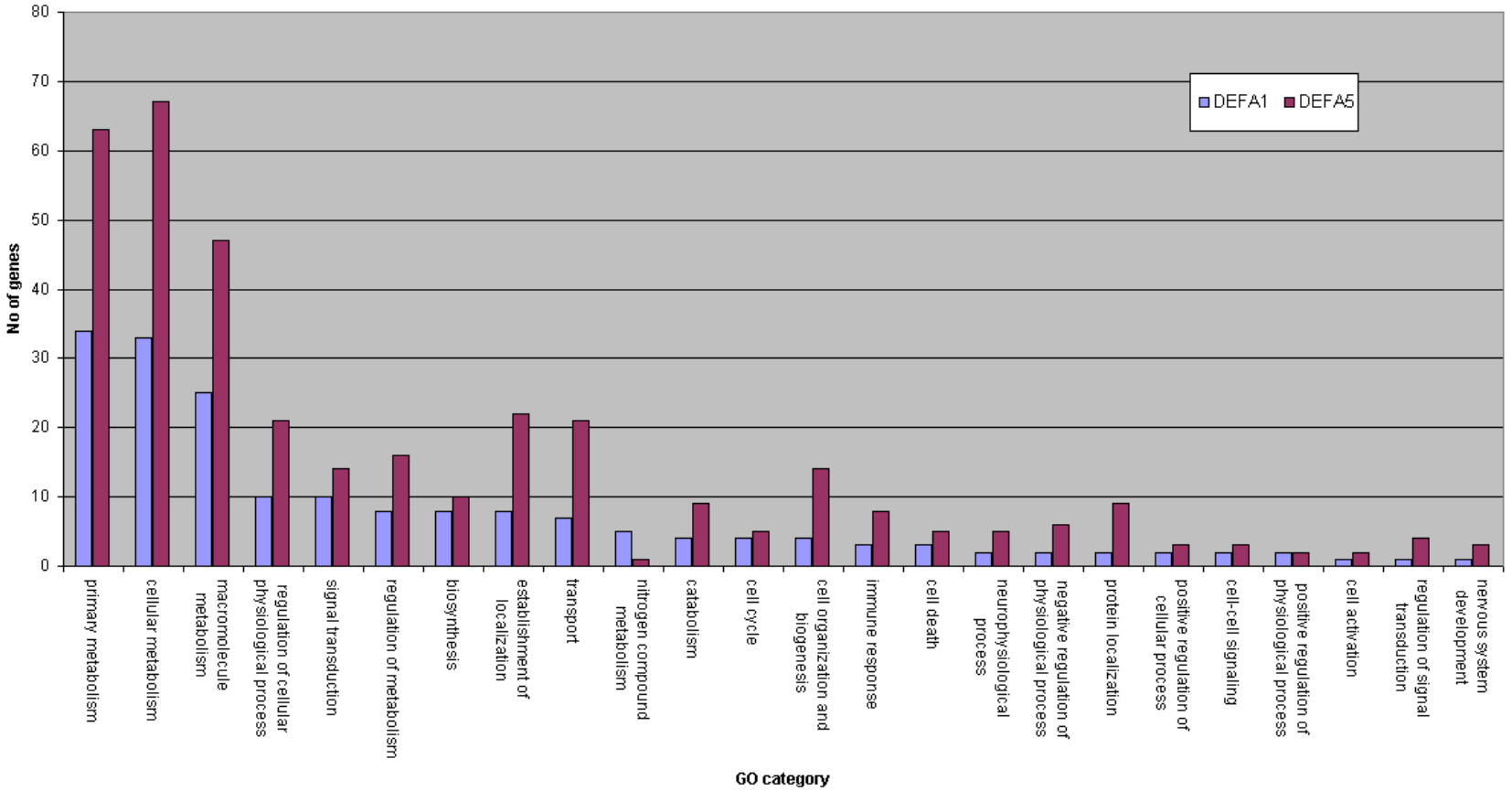


Figure 7.4: GO functions of DEFA5 gene hits that are exclusive to DEFA5 group

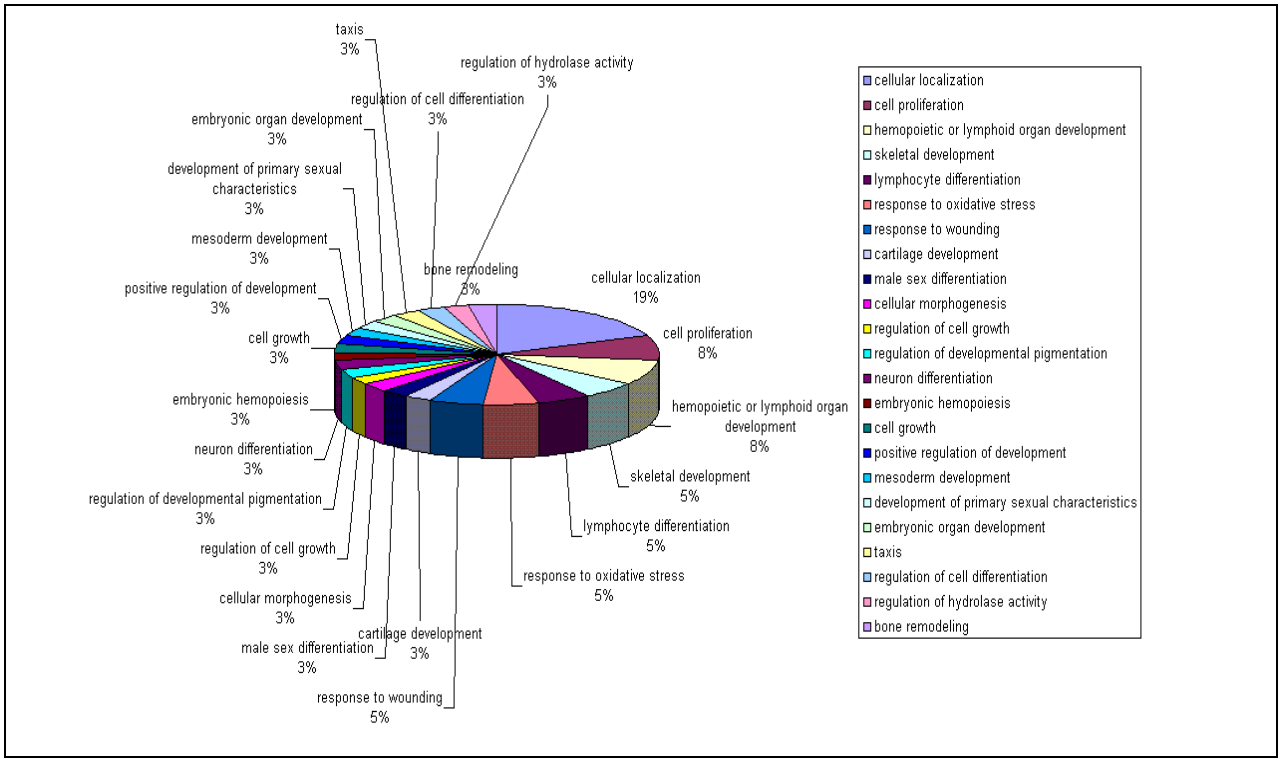


Figure 7.5: GO functions of DEFA1 gene hits that are exclusive to DEFA1 group

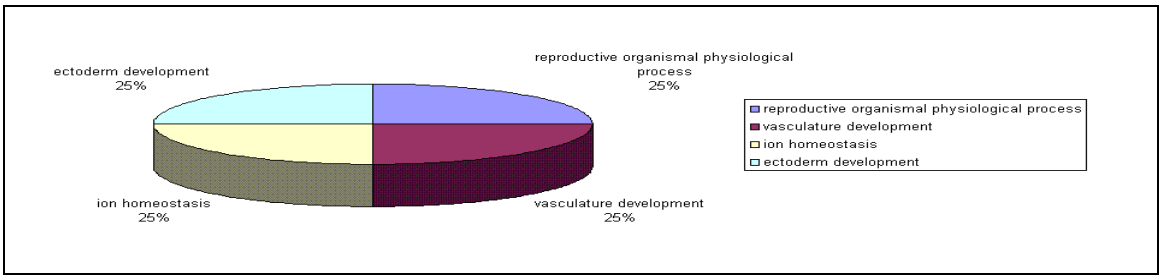


Table 7.4: Comparison of DEFA1 and DEFA5 gene hits based on pathways

The p-values are calculated by Fisher's exact test and multiple corrections have been done using the FDR procedure (Benjamini, 1995)

Common pathways between DEFA1 and DEFA5						
Pathways	**DEFA1 genes**	**No. of genes**	**DEFA5 genes**	**No of genes**	**Unadjusted pvalue**	**Adjusted pvalueFDR**
Focal adhesion	CAV2 CCND2	2	PARVA	1	5.55E-01	1
Cell cycle	CCND2	1	YWHAG CCNB	2	1	1
Prostaglandin and leukotriene metabolism	CBR1	1	KIAA0106	1	1	1
Wnt signaling pathway	CCND2	1	NMP238	1	1	1
Fructose and mannose metabolism	ALDR1	1	PFKL	1	1	1
Insulin signaling pathway	INS	1	PFKL MNK1	2	1	1
Starch and sucrose metabolism	DDX23	1	UGP2 NUDT5	2	1	1
Folate biosynthesis	DDX23	1	NUDT5	1	1	1
Purine metabolism	GUK1	1	PRIM1 NUDT5	2	1	1
Tyrosine metabolism	HRMT1L1	1	AOC1	1	1	1
Galactose metabolism	ALDR1	1	PFKL UGP2	2	1	1
Pentose and glucuronate interconversions	ALDR1	1	UGP2	1	1	1
Histidine metabolism	HRMT1L1	1	AOC1	1	1	1
Tryptophan metabolism	HRMT1L1	1	MID1 AOC1	2	1	1
MAPK signaling pathway	ARRB2 CASP5	2	MAP2K1IP1 MNK1	2	1	1
Glycine, serine and threonine metabolism	KBL	1	AOC1	1	1	1
DEFA1_specific pathway						
Pathways	**DEFA1 genes**	**No.of genes**	**DEFA5 genes**	**No. of genes**	**Unadjusted pvalue**	**Adjusted pvalueFDR**
Selenoamino acid metabolism	HRMT1L1 MARS	2	No genes	0	1.55E-01	1
Jak-STAT signaling pathway	PIAS2 CCND2	2	No genes	0	1.55E-01	1
Regulation of actin cytoskeleton	INS	1	No genes	0	4.00E-01	1
Methionine metabolism	MARS	1	No genes	0	4.00E-01	1
Dentatorubropallidoluysian atrophy (DRPLA)	INS	1	No genes	0	4.00E-01	1
N-Glycan biosynthesis	DPM1	1	No genes	0	4.00E-01	1
Pyruvate metabolism	ALDR1	1	No genes	0	4.00E-01	1
Oxidative phosphorylation	NDUFS5	1	No genes	0	4.00E-01	1
N-Glycan degradation	KIAA0935	1	No genes	0	4.00E-01	1
Nitrobenzene degradation	HRMT1L1	1	No genes	0	4.00E-01	1
Aminoacyl-tRNA synthetases	MARS	1	No genes	0	4.00E-01	1

Glycerolipid metabolism	ALDR1	1	No genes	0	4.00E-01	1
Aminosugars metabolism	KIAA0060	1	No genes	0	4.00E-01	1
Aminophosphonate metabolism	HRMT1L1	1	No genes	0	4.00E-01	1
Androgen and estrogen metabolism	HRMT1L1	1	No genes	0	4.00E-01	1
Fatty acid metabolism	PECI	1	No genes	0	4.00E-01	1
DEFA5_specific pathway						
Pathways	**DEFA1 genes**	**No. of genes**	**DEFA5 genes**	**No. of genes**	**Unadjusted pvalue**	**Adjusted pvalueFDR**
Alkaloid biosynthesis II	No genes	0	AOC1 KIAA0106	2	5.09E-01	1
Pyrimidine metabolism	No genes	0	TXNRD1 PRIM1	2	5.09E-01	1
Cell adhesion molecules (CAMs)	No genes	0	HLA-DMA CLDN2	2	5.09E-01	1
Adherens junction	No genes	0	SNAI1 SLUG	2	5.09E-01	1
Phenylalanine metabolism	No genes	0	AOC1 KIAA0106	2	5.09E-01	1
Calcium signaling pathway	No genes	0	HER3 CCNB	2	5.09E-01	1
Prion disease	No genes	0	GFAP	1	1	1
Neurodegenerative Disorders	No genes	0	GFAP	1	1	1
Proteasome	No genes	0	PSMA7	1	1	1
Methane metabolism	No genes	0	KIAA0106	1	1	1
Glycolysis / Gluconeogenesis	No genes	0	PFKL	1	1	1
Bile acid biosynthesis	No genes	0	NPD002	1	1	1
Stilbene, coumarine and lignin biosynthesis	No genes	0	KIAA0106	1	1	1
Nucleotide sugars metabolism	No genes	0	UGP2	1	1	1
1- and 2-Methylnaphthalene degradation	No genes	0	NPD002	1	1	1
Glycosylphosphatidylinositol(GPI)-anchor biosynthe	No genes	0	PIGT	1	1	1
2,4-Dichlorobenzoate degradation	No genes	0	KIAA0106	1	1	1
Valine, leucine and isoleucine degradation	No genes	0	AUH	1	1	1
Ribosome	No genes	0	RPS27L	1	1	1
DNA polymerase	No genes	0	PRIM1	1	1	1
beta-Alanine metabolism	No genes	0	AOC1	1	1	1
Tight junction	No genes	0	CLDN2	1	1	1
Pentose phosphate pathway	No genes	0	PFKL	1	1	1
Protein export	No genes	0	SPC18	1	1	1
Butanoate metabolism	No genes	0	KIAA0106	1	1	1
O-Glycan biosynthesis	No genes	0	WBSCR17	1	1	1
Arginine and proline metabolism	No genes	0	AOC1	1	1	1

7.5 Discussion

Using promoter models of alpha defensin, the human promoter dataset was scanned. Several of the predicted model hits matched with experimental co-expressed gene data.

The selection of genes from the promoter data set by the promoter models is dependent on the specificity of the motifs that are contained in the promoter model and the threshold that is used to scan the data set

The caveats of this work is that it had limited gene expression data to analyze which cover the various stimuli for studying co-expressed genes for parent AMP genes. Secondly, the promoter data set was limited since it did not cover promoter regions for the entire human genome. This is because the choice of TSS position was determined using strict rules (section 7.3.1), which greatly decreased the possibility of false TSSs, but the number of promoter sequences extracted was also decreased.

CX3CL1 is one of the predicted genes that came up for both the DEFA1 and DEFA5 model scan. CX3CL1 is one of the chemotaxins that stimulate NK cells (Morris and Ley, 2004). NK cells are known to produce alpha defensins and are directly involved in protection against microorganisms (Chalifour *et al.*, 2004). CKLFSF6 is another interesting gene hit that emerged from DEFA5 scan. This recently discovered gene that belongs to a novel gene family also indicates that chemotaxins are probably co-regulated along with genes of innate immunity.

The author found immune defense related genes for both DEFA1 and DEFA5 scans. DEFA1 hits were FKBP12, PSMB8, and DEFA5 hits were HLA-DMA, ILF2, G10P1, TTF, IFI30, AIF1, DHLAG. These genes are known to be co-expressed with

DEFA1, DEFA5 respectively. These findings corroborate the strategy that the author has used to find co-regulated genes for the AMP genes of interest.

The author compared the gene hits of DEFA1 and DEFA5 and found that both groups have a significant number of genes involved in metabolic pathways, signal transduction, biosynthesis, transport, cell death besides immune response. This supports the hypothesis that AMPs are involved in other pathways besides immune related ones. This analysis also indicated that though DEFA1 and DEFA5 belong to the same AMP family and have similarities in their transcription regulatory pathways, they still have some differences in terms of the different pathways they maybe involved in. However this statement may not be conclusive as the analysis was done based on GO terms and pathways which limits the number of genes that are taken consideration from the original pool of gene hits due to lack of annotation.

This analysis elucidate novel gene which are potentially co-regulated with the alpha defensins. This claim is backed up by the fact that these gene hits share the same promoter model with the alpha defensin genes. Furthermore, the gene hits have been subjected to different types of analyses based on GO term classification, expression data comparison, common transcription factors and regulators.

Though these results are just a drop in the sea of latent knowledge of transcriptional regulatory pathways for AMPs, it gives a glimpse of the complex interplay of different pathways and genes that could possibly be involved in influencing innate immunity and vice-versa.

.

7.6 Conclusion

The objective of this study was to find co-regulated genes for the two different alpha defensin genes, which have different sites of expression. To do this, the regulatory

elements in the promoter regions of the alpha defensin genes were searched and TFBS modules were found. These modules were then used to scan human promoter dataset to find potentially co-regulated genes. Promoter regions were analyzed for alpha defensin 1/alpha-defensin 3 and alpha-defensin 5 genes in human and mouse orthologs using *ab-initio* motif searching algorithm.

Seventeen predicted hits from alpha-defensin 1 promoter model scan and 177 predicted hits from alpha-defensin 5 coincided with experimentally found co-expressed genes with DEFA1, DEFA3 and DEFA5 genes.

The scan results reported gene hits CX3CL1 (chemokine (C-X3-C motif) ligand 1), an immunophilin FKBP12 (FK506-binding protein 1A), VSIG2 which is a member of the immunoglobulin domain cell adhesion molecule (cam), INS (insulin), PSMB8 which is involved in the process of antigen presentation to promoter model of DEFA1. Gene hits from DEFA5 promoter model HLA-DMA, ILF2, G10P1, TTF, IFI30, AIF1, DHLAG are immune genes. CX3CL1 comes up in scan for both DEFA1 and DEFA5 promoter model. CKLFSF6 is a gene that belongs to a novel superfamily of chemokine-like factor that emerged as a DEFA5 gene hit.

All the gene hits for DEFA1, DEFA5 emerged due to complete match of all motifs that made the promoter modules. However, not all gene hits matched experimentally found co-expressed genes. Comparison of functions of gene hits and the co-expressed genes showed significant similarity. This could probably indicate that these gene hits maybe co-regulated with DEFA1, DEFA5 respectively.

Part IV: Chapter 8 Discussion and Conclusion

It is hard to fail, but it is worse never to have tried to succeed.
(Theodore Roosevelt)

Computational biology appeared as a specialized discipline in the last quarter of the 20th century, and it is revolutionizing how biological research is conducted. Researchers are increasingly conducting searches in public databases for characterized sequences that match theirs before doing experiments to determine their function. Bioinformatics narrows down the number of essential experiments needed and thus expedites the discovery process. Technological innovations in biology have made possible the genome sequencing of various organisms, which has generated a tremendous amount of sequence data deposited in the databases. Inferring knowledge from these data has become a priority. Computational biology facilitates extraction of knowledge by high throughput analysis of vast chunks of data, which is not otherwise possible simply by experiments. This thesis exemplifies this statement by showing the usage of various computational methods to derive new knowledge from the antimicrobial peptide dataset. The main emphasis of this study has been to analyze the promoter region of AMPs and deduce promoter elements. However, in due course of the study many other interesting and novel results have been generated at both peptide and genomic level of AMPs.

The following paragraphs summarize the results of this thesis and discuss the implications of the findings.

8.1 Database of antimicrobial peptides

Databases serve as valuable resource for exploration of antimicrobial peptides, allowing users to query complex biological questions that may usually involve searching multiple sources. . In this thesis, the author created a publicly accessible database of antimicrobial peptides called ‘ANTIMIC’. ANTIMIC contains 1788 entries from both eukaryotic and prokaryotic origin. The process of creation of the database consisted of systematic data

collection, curation and cleaning that has been documented in Chapter 3 and can be reproduced if needed to update and enrich the database further. During the process of data collection and cleaning the public database records, some errors in the data were identified and corrected. Examples of errors include high redundancy due to maintenance of the same sequence in different public databases, discrepancies in primary sequences and conflicting annotation. Data checking and correction are thus critical for the improvement of data quality. Interpretation of unclean data is normally inaccurate and errors will be propagated in subsequent analysis where high data quality is important for accurate predictions. The creation of the ANTIMIC database was the first step towards a systematic sequence analysis of antimicrobial peptides. It helped in collating and systematizing the scattered information of antimicrobial peptides in one place and in easy access of information, which would help in the analysis process.

Data classification usually is the next step computational step applied to the data that facilitates better understanding of the data and sets ground for prediction work. There are two principal approaches taken for data classification, manual and automatic. Manual classifications are based on human expertise, facilitated by bioinformatic analyses, to cluster data into particular groups that share common properties defined by domain experts. Examples include the manually curated Swiss-Prot and PROSITE databases. The other approach is automatic classifications that depend on algorithms or models. Examples of automatic classification algorithms include self-organized maps, artificial neural network, and support vector machines which belong to the fields of artificial intelligence and machine learning ProDom (Kapetanovic *et al.*, 2004), and DOMO (Gracy and Argos, 1998), Pfam among others, address classification more systematically

with automated processes that classify entire protein sequence databases. The advantage of manual classification is the high quality of clustering but the final classification result may be irreproducible because of differences in the experts' knowledge. In contrast, automation is fully reproducible because of fixed rules written in computer programs and scalable to large data set, but implies caveat that the same threshold is used in the process and this may not be the optimal choice, thus potentially opening a way for the propagation of errors caused in this manner.

Multiple alignments of protein sequences are an effective way of classifying and also identifying conserved amino acids that provide clues to functional relationships among proteins. The patterns of amino acid variability in multiple sequence alignments reveal evolutionary pressure, mutation, recombination and genetic drift that spans millions of years Valdar, 2002. Conserved residues could be critical to the structure and function of a peptide. However, multiple alignments alone can go as far as aligning multiple sequences and indicating the conserved residues. To be able to increase the applicability of multiple alignments, HMMs are introduced. HMM approach models expectations of what unknown members of a protein family could be through the use of probabilities calculated from multiple alignments and assuming independence (except within consecutive deletions and insertions) among amino acids of a protein. Thus, each position is modeled separately; the concatenation of these amino acid probabilistic models is the protein model (Amitai, 1998). The author, implemented the HMM based machine learning approach to create a score based method of classification of AMPs that can be used to query for new AMPs and also predict the classification of new AMPs into known or new AMP families. HMM profiles or "signatures" were created for AMPs

based on prior knowledge of the AMP families. These profiles also proved to be useful in tagging of conserved residues that could potentially be important for the antimicrobial function for a particular family of AMP. The HMM based software was integrated into the ANTIMIC database as the ANTIMIC profile module.

8.2 Comparative genomic analysis of AMPs to find transcriptional regulatory elements

Understanding regulation of a gene or gene family at transcription level in recent years has gained momentum due to high-throughput genome sequencing of whole genomes and experimental techniques that have made it possible to explore non-coding regions. This understanding can be expedited through computational methods. Comparative genomics has long held the promise for the identification of response elements in eukaryotic genomes (Hardison *et al.*, 1997). Initially, searches for regulatory elements were conducted with consensus sequences and positional weight matrices and were confined to the detection of known elements. Now, *ab-initio* approaches show great potential for the identification of response elements in eukaryotic organisms (Lawrence *et al.*, 1993, Roth *et al.*, 1998). *Ab-initio* approach on regulatory region of multiple species and validation with co-expression information from DNA expression analysis experiments brings a powerful way to determine new regulatory elements. This approach allows for elucidation of new promoter elements (TFB motifs) which are previously unknown. This work shows implementation of this approach in understanding antimicrobial peptides and the author was able to find new insights into the transcriptional regulation of AMPs.

In order to collate the promoter sequences of various AMP genes, the author started with identification of AMP-coding cDNAs in the FANTOM3 data set and their

orthologous human or rat sequences. TBLASTN search was done on the FANTOM3 dataset using the ANTIMIC sequences and some additional AMP sequences from GenBank. The author was able to find 103 mouse AMP members that were new in FANTOM3. The sequences belonged to 28 families (alpha-defensin, alpha2casein, apoa2, beta-defensin, spag11, bpi, calgranulin, cathelicidin, cathepsinG, dbi, slpi, enhancer of rudimentary homolog, granulin, hepcidin, histone2a, IFN-inducible antiviral protein Mx, lactoferrin, lysozyme, mbp, melanotropin alpha, ovotransferrin, proenkephalin 1, sap2, secretogranin, skiv2l, spyy, vasostatin, vip and zap).

The extraction of promoter sequences of AMPs involved a systematic collection from various sources. Mouse promoter sequence extraction involved sequential steps of finding the TSS location that was determined by using the start position of the first exon of the FANTOM cDNA-genome mapping data. Upstream region of 1000 base pairs and downstream region of 200 base pairs were then extracted by mapping the TSS location to the mouse genome data from UCSC.

The promoter analysis was done on genes from 22 AMP families. The promoter regions of AMPcgs in the three species (human, mouse and rat) were screened for motifs by an *ab-initio* motif finding method and analyzed for promoter characteristics (TFBS motifs). Many of the motifs that were detected are known from previous experimental studies to be involved in control of AMPs (Chapter 5). This corroborates the computational method used to detect TFBS motifs. The analysis showed that the key transcriptional regulators are likely to be TFs of the liver-, nervous system-specific and NHR group. Nuclear hormone receptors (NHRs) were prominent among the core TF group. NHR such as GR, RXR-alpha, AR, VDR, T3R-alpha and RAR-alpha1 emerged as

the frequently occurring NHR, some of which are implicated in immune responses. Non-NHR TFs such as Meis1a and Meis1b, Sp1, NF-1, AP-2 and c-Myb were the other TFs found in various AMP families. These TF groups consist of transcription regulators that are involved in diverse physiological functions, including control of embryonic development, cell differentiation and homeostasis, but also in immune response. This reiterates that AMPs are involved in pathways other than innate immunity.

The analyses of the promoter regions of AMPs lead to several other interesting observations. Analysis of alpha defensin promoter regions showed that the conservation of the motifs across different species correlated to the phylogenetic groupings that have been studied previously by other groups for alpha defensins (Chapter 5). It was observed that the motif combinations that are shared between myeloid and enteric specific alpha defensins largely differ between rodents and primates. The rat Defa and enteric-expressed mouse defcr2 promoter regions share the motifs 20 (AR PXR-1: RXR-alpha) –7 (POU1F1a, POU2F1) – 4 (RAR-alpha1, RXR-alpha) (20-7-4) arrangement (Chapter 6). In contrast, the primate myeloid–expressed (Hosa_DEFA4, Patr_DEFA4, Hosa_DEFA3) and enteric-expressed (Hosa_DEFA5 and Patr_DEFA5) alpha defensins share the motif organization (20-10-11-19) (Chapter 5). This small study acts as an example of investigating phylogeny with respect to regulatory content analysis. It initiates a new perspective into the study of evolution of various AMP gene families to find out about their common ancestry and divergence and intrigues further investigation on whether a classification maybe possible based on regulatory regions and gene structure for AMP genes.

Three potential TF-binding motifs that were enriched in promoters of AMPcgs are novel. This sets precedence for experimental validation of these cis-elements. (Chapter 5). Another four motifs were found to be species-specific or lineage-specfic in the context of regulation of individual AMPcg families. (Chapter 5).

The next step was to look at cohorts of putative TFBS motifs and deduce a common framework or model of TFBS. The author generated promoter models for PENK, DEFA5, DEFA1 and ZAP AMP genes (Chapter 6). Most of these models consisted of known and novel TFBSs. This has been the first attempt to generate promoter models for AMP families.

The significance of these models was in their usage to be able to extract co-regulated genes that share the same promoter models. This was demonstrated with the alpha defensin promoter models (DEFA1, DEFA5). The promoter models (alpha defensins) were used to scan the human promoter dataset (Chapter 7). The scanned hits found using the promoter models coincided with known co-expressed genes of the parent AMPs (Chapter 7) and this vindicates the computational approach used. Many novel gene hits emerged in the promoter model scan that had similar GO based categorization as the experimentally known co-expressed gene group (Chapter 7) indicating the possibility that these genes can also be co-expressed under different conditions with AMPs. CX3CL1 is one of the predicted genes that came up for both the DEFA1 and DEFA5 model scan which is not known to be experimentally co-expressed with alpha defensins. However, its function as a chemotaxin and involvement in adaptive immune response indicates that it maybe possibly co-expressed with alpha defensins. Results also

show that the scanned gene hits for alpha defensins are also regulated by immune pathway related TFs such as NF-kappaB, GATA-4 and Evi-1 (Chapter 7). From the comparison of DEFA1 and DEFA5 gene hits it can be extrapolated that though there are many common pathways such as cell cycle, MAPK signaling pathway, insulin signaling pathway, etc.) in which both gene groups are involved, there are also different pathways which appear exclusive to only one gene group (either DEFA1 or DEFA5) such as methionine metabolism, androgen and estrogen metabolism, selenoamino acid metabolism which showed up only for DEFA1 gene hits. etc. (Chapter 7). Glycolysis, pyrimidine metabolism, alkaloid biosynthesis II etc were some pathways that were observed only for the DEFA5 genes. Therefore, this is perhaps an indication that though DEFA1 and DEFA5 belong to the same AMP family, they are involved in different gene networks. The methodology of promoter element analysis is applicable to any multigene families with diverse functions (i.e. cytokines and chemokine ligands and receptors) and more importantly for establishing basic functional assignments for transcripts with unknown functions.

In summary, this analysis shows that AMPs from different families have multiple roles in cells, which implies that they are likely to be regulated in such a manner that they can fulfill their roles. This means that we should expect great variability in their promoters. However, since they also have some common roles in immune response, we also expect that part of their promoter characteristics could be similar.

A great part of the author's computational findings fit into the current knowledge about regulation of AMPs and also generate new hypotheses that await experimental validation. This study acts as a paradigm for the use of computational tools such as DMB , module generation program etc. to understand parts of regulatory regions for any set of genes and collation of the data to fit a bigger network of the underlying workings of the transcriptional regulation.

This thesis unveils opputunities for development and expansion of research in the area of transcriptional regulation of AMPs and also other gene families.

Two principal directions for the development of bioinformatics in the field of antimicrobial peptides and transcriptional regulation, are namely the development of databases and computational analysis. The development of database includes data update on top of data integration, data cleaning and integration of bioinformatic tools. Data update focuses on adding new antimicrobial peptide sequences identified and new 3D-structures solved. The author suggests that the ANTIMIC database can be enriched further with additional annotation to make it a comprehensive and composite repository. It can contain in addition to the current version additional information of the known gene structures, promoter regions, transcription factors that have been published in literature for antimicrobial peptides and also computationally found into the data warehouse. The new data can help to verify hypotheses made during analyses of initial dataset while new information can provide insights for further analysis.

Part IV: Chapter 9: Future work

Years teach us more than books.
(Berthold Auerbach)

9.1 Experimental work

In recent years, there has been a rapid increase in our knowledge of understanding the role of regulatory regions as "switches" of various pathways and onset of diseases. Both experimental and computational tools have improved over the years to enable such a growth in our understanding. Now it is very much feasible to deduce missing links in pathways by application of a combination of experimental and computational techniques. For example, a study reported the use of promoter analysis to identify novel genes showing functional relevance in cell proliferation in a colon cancer model. The analysis yielded some known proliferation-associated genes, such as HERG1 and MCM7, and a number of genes not previously implicated in cell proliferation in cancer, such as TSPAN3, Necdin and APLP2. Suppression of TSPAN3 and APLP2 by siRNA was performed and confirmed by RT-PCR. It was seen that inhibition of these genes significantly inhibited cell proliferation in colon cancer cell line (Moss AC, 2007).

In another example, promoter modeling was applied to link disease-associated genes to potential regulatory networks. This approach was applied to a Maturity Onset Diabetes of the Young (MODY)-associated gene list, which yielded two models connecting functionally interacting genes within MODY-related insulin/glucose signaling pathways (Dohr *et al.*, 2005).

Another group has identified some novel potential transcriptional regulators and pathways involved at different stages of spermatogenesis based on bioinformatic and promoter analysis. The analysis was done on SAGE data on the transcriptome of mouse type A spermatogonia (Spga), pachytene spermatocytes (Spcy), and round spermatids (Sptd) (Lee *et al.*, 2006).

In this thesis as well, a number of hypotheses has been generated that have good concordance with some of the existing knowledge in the field of AMPs and innate immunity. However, the computationally inferred hypotheses can only be tested in experiments. This section discusses the novel findings obtained from the analyses and the experimental approaches to validate them.

1. Several of the NHR group TFs such as RXR-alpha, AR, T3R-alpha, RAR-alpha, LXR-alpha:RXR-alpha appear as the frequently occurring candidates in AMPS. Their presence in the promoter regions requires validation.

2. VDR and GR have been reported in scientific literature to be regulating expression of some AMP genes such as beta defensins. Our analysis indicates their presence in many other AMP genes.

3. NF-1 and NKX2-1 which have not yet been implicated to be involved immunomodulatory pathways have appeared frequently in many AMP genes in the analysis.

4. Transcriptional regulator *c-myb* is involved in lymphocyte development. Experimental validation of its functional binding site in atleast an AMP gene can indicate its direct involvement in innate immunity.

The author suggests genome-scale location analysis (Ren *et al.*, 2000) followed by chromatin immunoprecipitation (ChIP) can identify promoters bound by the computationally predicted TFs in various tissue cells. These experiments are expected to

clarify which promoters and TFs are specific for certain tissue cells and how many AMPcgs are regulated by a TF, TF pair or multiple TFs.

This study has also elucidated genes that are novel candidates for co-regulation with AMPs such as CX3CL1. Microarray experiments that are pathway specific or ligand specific stimulated such as LPS and TNF can be carried out to validate co-expression of the candidate genes with certain AMP genes.

Another experimental approach could be knocking out of a particular AMP gene and checking for the effects of the deletion on a pathway of interest from the suggested pathways in this study. This can facilitate validation of a involvement of a particular AMP gene in regulation of the pathway. Eventually, the combination of both computational and experimental will facilitate construction of mechanistic models of AMPcg regulatory transcription networks.

9.2 Computational work

Expression arrays yield high dimensional data that facilitates the deduction of temporal and special activation of groups of genes. Through this technology it has been possible to find the differential gene expression patterns in normal and diseased tissues as well as the response of tissues to the application of therapeutic reagents.

Computational analysis on gene expression data such as application of clustering algorithms facilitate elucidation of co-regulated set of genes. Information about underlying transcriptional regulatory networks responsible for the observed expression patterns is not directly deductible based on cDNA sequenes used to generate the arrays.

Regulation of expression is determined to a large extent by the promoter sequences of the individual genes (and/or enhancers). The availability of the complete human genome sequence now provides the molecular basis for the identification of many regulatory regions. Promoter sequences for specific cDNAs can be obtained reliably from genomic sequences by exon mapping or by mapping full length cDNAs. A sufficient pool of promoter sequences can allow deduction to candidates in a network solely on bioinformatic analysis as has been demonstrated in this thesis. Generation of promoter models based on comparative promoter analysis of co-regulated genes and groups of genes leads the way to understanding regulatory networks. Such modules represent the molecular mechanisms through which regulatory networks influence gene expression. This approach also provides a powerful alternative for elucidating the functional features of genes with no detectable sequence similarity, by linking them to other genes on the basis of their common promoter structures.

As a part of future work, the author thus proposes promoter model scanning across whole genomes (for eg. human, mouse) to find putatively new AMP related sequences that can not be extracted by BLAST or other sequence similarity tools due to short length of the exon or low similarity.

Another possibility is for generation of promoter models for other AMP families that have not been addressed in this thesis for example beta-defensins and promoter scan with the models generated. This will enable expansion of our knowledge in the realm of regulatory networks in which AMP genes are involved.

It would also be interesting to do the scan across other available promoter datasets like mouse and rat and do a comparative genomic study of the gene hits for a particular AMP model to find how similarities and differences among them.

In this thesis, the author restricted the findings of regulatory elements in the promoter region. The same computational approach can be extended to find the regulatory elements in intergenic regions and 3'UTR region. This would lead to novel findings in these regulatory regions of which we have limited knowledge in current time.

References

Life is either a daring adventure or nothing.
(Helen Keller)

Aarbiou, J., Ertmann, M.*, et al.*,2002. Human neutrophil defensins induce lung epithelial cell proliferation in vitro. J Leukoc Biol 72, 167-174.

Aarbiou, J., Verhoosel, R. M.*, et al.*,2004. Neutrophil defensins enhance lung epithelial wound closure and mucin gene expression in vitro. Am J Respir Cell Mol Biol 30, 193-201.

Adamietz, P., Bredehorst, R.*, et al.*,1978. ADP-ribosylated histone H1 from HeLa cultures. Fundamental differences to (ADP-ribose)n-histone H1 conjugates formed into vitro. Eur J Biochem 91, 317-326.

Albig, W., Trappe, R.*, et al.*,1999. The human H2A and H2B histone gene complement. Biol Chem 380, 7-18.

Alkema, W. B., Johansson, O.*, et al.*,2004. MSCAN: identification of functional clusters of transcription factor binding sites. Nucleic Acids Res 32, W195-198.

Al-Shahrour, F., Diaz-Uriarte, R.*, et al.*,2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20, 578-580.

Al-Shahrour, F., Minguez, P.*, et al.*,2006. BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Res 34, W472-476.

Altschul, S. F., Gish, W.*, et al.*,1990. Basic local alignment search tool. J Mol Biol 215, 403-410.

Altschul, S. F., Madden, T. L.*, et al.*,1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.

Amitai, M.,1998. Hidden models in biopolymers. Science 282, 1436-1437.

Arpin, C., Pihlgren, M., *et al.*,2000. Effects of T3R alpha 1 and T3R alpha 2 gene deletion on T and B lymphocyte development. J Immunol 164, 152-160.

Autieri, M. V., Carbone, C., *et al.*,2000. Expression of allograft inflammatory factor-1 is a marker of activated human vascular smooth muscle cells and arterial injury. Arterioscler Thromb Vasc Biol 20, 1737-1744.

Avellar, M. C., Honda, L., *et al.*,2004. Differential expression and antibacterial activity of epididymis protein 2 isoforms in the male reproductive tract of human and rhesus monkey (Macaca mulatta). Biol Reprod 71, 1453-1460.

Baeuerle, P. A. and Henkel, T.,1994. Function and activation of NF-kappa B in the immune system. Annu Rev Immunol 12, 141-179.

Bailey, T. L. and Elkan, C.,1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2, 28-36.

Bailey, T. L. and Elkan, C.,1995. The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol 3, 21-29.

Bailey, T. L. and Noble, W. S.,2003. Searching for statistically significant regulatory modules. Bioinformatics 19 Suppl 2, II16-II25.

Bairoch, A., Boeckmann, B., *et al.*,2004. Swiss-Prot: juggling between evolution and stability. Brief Bioinform 5, 39-55.

Bajic, V. B. and Seah, S. H.,2003. Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes. Nucleic Acids Res 31, 3560-3563.

Bajic, V. B., Seah, S. H., *et al.*,2002. Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. Bioinformatics 18, 198-199.

Bajic, V. B., Tan, S. L., *et al.*,2006. Mice and men: their promoter properties. PLoS Genet 2, e54.

Bajic, V. B., Tan, S. L., *et al.*,2004. Promoter prediction analysis on the whole human genome. Nat Biotechnol 22, 1467-1473.

Ballow, M., Wang, X., *et al.*,2003. Expression and regulation of nuclear retinoic acid receptors in human lymphoid cells. J Clin Immunol 23, 46-54.

Bals, R., Weiner, D. J., *et al.*,1999. Augmentation of innate host defense by expression of a cathelicidin antimicrobial peptide. Infect Immun 67, 6084-6089.

Banno, T., Gazel, A., *et al.*,2004. Effects of tumor necrosis factor-alpha (TNF alpha) in epidermal keratinocytes revealed using global transcriptional profiling. J Biol Chem 279, 32633-32642.

Barrett, C., Hughey, R., *et al.*,1997. Scoring hidden Markov models. Comput Appl Biosci 13, 191-199.

Bastian, A. and Schafer, H.,2001. Human alpha-defensin 1 (HNP-1) inhibits adenoviral infection in vitro. Regul Pept 101, 157-161.

Bateman, A., Birney, E., *et al.*,1999. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 27, 260-262.

Baveye, S., Elass, E., *et al.*,1999. Lactoferrin: a multifunctional glycoprotein involved in the modulation of the inflammatory process. Clin Chem Lab Med 37, 281-286.

Bechinger, B., Zasloff, M., *et al.*,1993. Structure and orientation of the antibiotic peptide magainin in membranes by solid-state nuclear magnetic resonance spectroscopy. Protein Sci 2, 2077-2084.

Bellamy, W., Takase, M., *et al.*,1992. Antibacterial spectrum of lactoferricin B, a potent bactericidal peptide derived from the N-terminal region of bovine lactoferrin. J Appl Bacteriol 73, 472-479.

Bellamy, W., Wakabayashi, H., *et al.*,1993. Killing of Candida albicans by lactoferricin B, a potent antimicrobial peptide derived from the N-terminal region of bovine lactoferrin. Med Microbiol Immunol (Berl) 182, 97-105.

Benjamini, Y. a. H., Y.,1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57, 289–300.

Benson, D. A., Karsch-Mizrachi, I., *et al.*,2004. GenBank: update. Nucleic Acids Res 32 Database issue, D23-26.

Benson, D. A., Karsch-Mizrachi, I., *et al.*,2005. GenBank. Nucleic Acids Res 33 Database Issue, D34-38.

Birney, E., Andrews, T. D., *et al.*,2004. An overview of Ensembl. Genome Res 14, 925-928.

Blondelle, S. E., Simpkins, L. R., *et al.*,1993. Influence of tryptophan residues on melittin's hemolytic activity. Biochim Biophys Acta 1202, 331-336.

Boman, H. G.,2000. Innate immunity and the normal microflora. Immunol Rev 173, 5-16.

Boman, H. G., Agerberth, B., *et al.*,1993. Mechanisms of action on Escherichia coli of cecropin P1 and PR-39, two antibacterial peptides from pig intestine. Infect Immun 61, 2978-2984.

Brahmachary, M., Krishnan, S. P., *et al.*,2004. ANTIMIC: a database of antimicrobial sequences. Nucleic Acids Res 32, D586-589.

Brotz, H., Josten, M., *et al.*,1998. Role of lipid-bound peptidoglycan precursors in the formation of pores by nisin, epidermin and other lantibiotics. Mol Microbiol 30, 317-327.

Brown, N. P., Leroy, C., *et al.*,1998. MView: a web-compatible database search or multiple alignment viewer. Bioinformatics 14, 380-381.

Bulet, P., Dimarcq, J. L., *et al.*,1993. A novel inducible antibacterial peptide of Drosophila carries an O-glycosylated substitution. J Biol Chem 268, 14893-14897.

Butler, J. E. and Kadonaga, J. T.,2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev 16, 2583-2592.

Calvo, K. R., Knoepfler, P. S., *et al.*,2001. Meis1a suppresses differentiation by G-CSF and promotes proliferation by SCF: potential mechanisms of cooperativity with Hoxa9 in myeloid leukemia. Proc Natl Acad Sci U S A 98, 13120-13125.

Can, T., Camoglu, O., *et al.*,2004. Automated protein classification using consensus decision. Proc IEEE Comput Syst Bioinform Conf 224-235.

Cao, Z. A., Moore, B. B., *et al.*,2000. Identification of an IFN-gamma responsive region in an intron of the invariant chain gene. Eur J Immunol 30, 2604-2611.

Carlsson, A., Engstrom, P., *et al.*,1991. Attacin, an antibacterial protein from Hyalophora cecropia, inhibits synthesis of outer membrane proteins in Escherichia coli by interfering with omp gene transcription. Infect Immun 59, 3040-3045.

Carninci, P., Kasukawa, T., *et al.*,2005. The transcriptional landscape of the mammalian genome. Science 309, 1559-1563.

Chalifour, A., Jeannin, P., *et al.*,2004. Direct bacterial protein PAMP recognition by human NK cells involves TLRs and triggers alpha-defensin production. Blood 104, 1778-1783.

Chaly, Y. V., Paleolog, E. M., *et al.*,2000. Neutrophil alpha-defensin human neutrophil peptide modulates cytokine production in human monocytes and adhesion molecule expression in endothelial cells. Eur Cytokine Netw 11, 257-266.

Chang, T. L., Vargas, J., Jr., *et al.*,2005. Dual role of alpha-defensin-1 in anti-HIV-1 innate immunity. J Clin Invest 115, 765-773.

Chaves, M., Albert, R., *et al.*,2005. Robustness and fragility of Boolean models for genetic regulatory networks. J Theor Biol 235, 431-449.

Chen, H., Xu, Z., *et al.*,2006. Recent advances in the research and development of human defensins. Peptides 27, 931-940.

Chen, K. C., Wang, T. Y., *et al.*,2005. A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae. Bioinformatics 21, 2883-2890.

Chen, W. G., West, A. E., *et al.*,2003. Upstream stimulatory factors are mediators of Ca2+-responsive transcription in neurons. J Neurosci 23, 2572-2581.

Choe, Y. H., Oh, Y. J., *et al.*,2003. Lactoferrin sequestration and its contribution to iron-deficiency anemia in Helicobacter pylori-infected gastric mucosa. J Gastroenterol Hepatol 18, 980-985.

Chong, A., Zhang, G., *et al.*,2003. FIE2: A program for the extraction of genomic DNA sequences around the start and translation initiation site of human genes. Nucleic Acids Res 31, 3546-3553.

Chuen, C. K., Li, K., *et al.*,2004. Interleukin-1beta up-regulates the expression of thrombopoietin and transcription factors c-Jun, c-Fos, GATA-1, and NF-E2 in megakaryocytic cells. J Lab Clin Med 143, 75-88.

Chugh, J. K. and Wallace, B. A.,2001. Peptaibols: models for ion channels. Biochem Soc Trans 29, 565-570.

Cohen, C. D., Klingenhoff, A., *et al.*,2006. Comparative promoter analysis allows de novo identification of specialized cell junction-associated proteins. Proc Natl Acad Sci U S A 103, 5682-5687.

Comb, M. J., Kobierski, L., *et al.*,1992. Regulation of opioid gene expression: a model to understand neural plasticity. NIDA Res Monogr 126, 98-112.

Conover, W. J.,1998. Practical Nonparametric Statistics (3rd Ed.).

Courselaud, B., Pigeon, C., *et al.*,2002. C/EBPalpha regulates hepatic transcription of hepcidin, an antimicrobial peptide and regulator of iron metabolism. Cross-talk between C/EBP pathway and iron metabolism. J Biol Chem 277, 41163-41170.

Couto, M. A., Harwig, S. S., *et al.*,1993. Selective inhibition of microbial serine proteases by eNAP-2, an antimicrobial peptide from equine neutrophils. Infect Immun 61, 2991-2994.

Crovella, S., Antcheva, N., *et al.*,2005. Primate beta-defensins--structure, function and evolution. Curr Protein Pept Sci 6, 7-21.

Cunliffe, R. N.,2003. Alpha-defensins in the gastrointestinal tract. Mol Immunol 40, 463-467.

Cvekl, A., Sax, C. M.*, et al.*,1994. A complex array of positive and negative elements regulates the chicken alpha A-crystallin gene: involvement of Pax-6, USF, CREB and/or CREM, and AP-1 proteins. Mol Cell Biol 14, 7363-7376.

Dangl, J. L. and Jones, J. D.,2001. Plant pathogens and integrated defence responses to infection. Nature 411, 826-833.

Danilova, N.,2006. The evolution of immune mechanisms. J Exp Zoolog B Mol Dev Evol

Darveau, R. P., Cunningham, M. D.*, et al.*,1991. Beta-lactam antibiotics potentiate magainin 2 antimicrobial activity in vitro and in vivo. Antimicrob Agents Chemother 35, 1153-1159.

De Gregorio, E., Spellman, P. T.*, et al.*,2002. The Toll and Imd pathways are the major regulators of the immune response in Drosophila. Embo J 21, 2568-2579.

de Kruijff, B.,1990. Cholesterol as a target for toxins. Biosci Rep 10, 127-130.

de Leeuw, E., Burks, S. R.*, et al.*,2007. Structure-dependent functional properties of human defensin 5. FEBS Lett 581, 515-520.

De Vos, P., Schmitt, J.*, et al.*,1994. Human androgen receptor expressed in HeLa cells activates transcription in vitro. Nucleic Acids Res 22, 1161-1166.

Deen, P. M., Terwel, D.*, et al.*,1991. Structural analysis of the entire proopiomelanocortin gene of Xenopus laevis. Eur J Biochem 201, 129-137.

Deshpande, N., Addess, K. J.*, et al.*,2005. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. Nucleic Acids Res 33, D233-237.

Diamond, G., Kaiser, V., *et al.*,2000. Transcriptional regulation of beta-defensin gene expression in tracheal epithelial cells. Infect Immun 68, 113-119.

Dohr, S., Klingenhoff, A., *et al.*,2005. Linking disease-associated genes to regulatory networks via promoter organization. Nucleic Acids Res 33, 864-872.

Durham, P. L., Sharma, R. V., *et al.*,1997. Repression of the calcitonin gene-related peptide promoter by 5-HT1 receptor activation. J Neurosci 17, 9545-9553.

Durr, M. and Peschel, A.,2002. Chemokines meet defensins: the merging concepts of chemoattractants and antimicrobial peptides in host defense. Infect Immun 70, 6515-6517.

Dushay, M. S., Roethele, J. B., *et al.*,2000. Two attacin antibacterial genes of Drosophila melanogaster. Gene 246, 49-57.

E Huang, L. Y., R Chowdhary, A Kassim, VB Bajic,2005. An algorithm for ab initio DNA motif detection. Information Processing and Living Systems, World Scientific 611-614.

Eckmann, L.,2005. Defence molecules in intestinal innate immunity against bacterial infections. Curr Opin Gastroenterol 21, 147-151.

Eddy, S. R.,1995. Multiple alignment using hidden Markov models. Proc Int Conf Intell Syst Mol Biol 3, 114-120.

Eddy, S. R.,1998. Profile hidden Markov models. Bioinformatics 14, 755-763.

Edwards, P. A., Kennedy, M. A., *et al.*,2002. LXRs; oxysterol-activated nuclear receptors that regulate genes controlling lipid homeostasis. Vascul Pharmacol 38, 249-256.

Elholm, M., Bjerking, G., *et al.*,1996. Regulatory elements in the promoter region of the rat gene encoding the acyl-CoA-binding protein. Gene 173, 233-238.

Fang, X. M., Shu, Q., *et al.*,2003. Differential expression of alpha- and beta-defensins in human peripheral blood. Eur J Clin Invest 33, 82-87.

Farrar, W. L., Ferris, D. K., *et al.*,1989. The molecular basis of immune cytokine action. Crit Rev Ther Drug Carrier Syst 5, 229-261.

Fehlbaum, P., Rao, M., *et al.*,2000. An essential amino acid induces epithelial beta -defensin expression. Proc Natl Acad Sci U S A 97, 12723-12728.

Ferre, R., Badosa, E., *et al.*,2006. Inhibition of plant-pathogenic bacteria by short synthetic cecropin A-melittin hybrid peptides. Appl Environ Microbiol 72, 3302-3308.

Fessele, S., Boehlk, S., *et al.*,2001. Molecular and in silico characterization of a promoter module and C/EBP element that mediate LPS-induced RANTES/CCL5 expression in monocytic cells. Faseb J 15, 577-579.

Fisher, R. A.,1922. On the interpretation of ?2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85(1), 87-94.

Frech, K., Danescu-Mayer, J., *et al.*,1997. A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. J Mol Biol 270, 674-687.

Frith, M. C., Hansen, U., *et al.*,2004. Finding functional sequence elements by multiple local alignment. Nucleic Acids Res 32, 189-200.

Frith, M. C., Li, M. C., *et al.*,2003. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. Nucleic Acids Res 31, 3666-3668.

Frohm Nilsson, M., Sandstedt, B., *et al.*,1999. The human cationic antimicrobial protein (hCAP18), a peptide antibiotic, is widely expressed in human squamous epithelia and colocalizes with interleukin-6. Infect Immun 67, 2561-2566.

Froy, O.,2005. Regulation of mammalian defensin expression by Toll-like receptor-dependent and independent signalling pathways. Cell Microbiol 7, 1387-1397.

Frye, M., Bargon, J., *et al.*,2000. Expression of human alpha-defensin 5 (HD5) mRNA in nasal and bronchial epithelial cells. J Clin Pathol 53, 770-773.

Fu, W., Shah, S. R., *et al.*,1997. Transactivation of proenkephalin gene by HTLV-1 tax1 protein in glial cells: involvement of Fos/Jun complex at an AP-1 element in the proenkephalin gene promoter. J Neurovirol 3, 16-27.

Fujii, Y., Imanishi, T., *et al.*,2004. [H-Invitational Database: integrated database of human genes]. Tanpakushitsu Kakusan Koso 49, 1937-1943.

Fujiwara, K., Ochiai, M., *et al.*,2004. Global gene expression analysis of rat colon cancers induced by a food-borne carcinogen, 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine. Carcinogenesis 25, 1495-1505.

Ganz, T.,1987. Extracellular release of antimicrobial defensins by human polymorphonuclear leukocytes. Infect Immun 55, 568-571.

Ganz, T.,2003. Defensins: antimicrobial peptides of innate immunity. Nat Rev Immunol 3, 710-720.

Ganz, T., Selsted, M. E., *et al.*,1985. Defensins. Natural peptide antibiotics of human neutrophils. J Clin Invest 76, 1427-1435.

Gao, G., Guo, X., *et al.*,2002. Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein. Science 297, 1703-1706.

Garcia-Garcia, L., Harbuz, M. S., *et al.*,1998. RU-486 blocks stress-induced enhancement of proenkephalin gene expression in the paraventricular nucleus of rat hypothalamus. Brain Res 786, 215-218.

Ge, Y., MacDonald, D. L., *et al.*,1999. In vitro antibacterial properties of pexiganan, an analog of magainin. Antimicrob Agents Chemother 43, 782-788.

Giacometti, A., Cirioni, O., *et al.*,2000. In-vitro activity and killing effect of polycationic peptides on methicillin-resistant Staphylococcus aureus and interactions with clinically used antibiotics. Diagn Microbiol Infect Dis 38, 115-118.

Gordon, Y. J., Romanowski, E. G., *et al.*,2005. A review of antimicrobial peptides and their therapeutic potential as anti-infective drugs. Curr Eye Res 30, 505-515.

Gracy, J. and Argos, P.,1998. DOMO: a new database of aligned protein domains. Trends Biochem Sci 23, 495-497.

Gronostajski, R. M.,2000. Roles of the NFI/CTF gene family in transcription and development. Gene 249, 31-45.

Guex, N. and Peitsch, M. C.,1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18, 2714-2723.

Guo, X., Carroll, J. W., *et al.*,2004. The zinc finger antiviral protein directly binds to specific viral mRNAs through the CCCH zinc finger motifs. J Virol 78, 12781-12787.

Hahm, S. H. and Eiden, L. E.,1998. Cis-regulatory elements controlling basal and inducible VIP gene transcription. Ann N Y Acad Sci 865, 10-26.

Halees, A. S., Leyfer, D., *et al.*,2003. PromoSer: A large-scale mammalian promoter and transcription start site identification service. Nucleic Acids Res 31, 3554-3559.

Han, W., Ding, P., *et al.*,2003. Identification of eight genes encoding chemokine-like factor superfamily members 1-8 (CKLFSF1-8) by in silico cloning and experimental validation. Genomics 81, 609-617.

Hancock, R. E. and Diamond, G.,2000. The role of cationic antimicrobial peptides in innate host defences. Trends Microbiol 8, 402-410.

Hansen, H. O., Andreasen, P. H., *et al.*,1991. Induction of acyl-CoA-binding protein and its mRNA in 3T3-L1 cells by insulin during preadipocyte-to-adipocyte differentiation. Biochem J 277 (Pt 2), 341-344.

Harder, J., Meyer-Hoffert, U., *et al.*,2000. Mucoid Pseudomonas aeruginosa, TNF-alpha, and IL-1beta, but not IL-6, induce human beta-defensin-2 in respiratory epithelia. Am J Respir Cell Mol Biol 22, 714-721.

Hardison, R. C., Oeltjen, J., *et al.*,1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res 7, 959-966.

Hayashi, R., Wada, H., *et al.*,2004. Effects of glucocorticoids on gene transcription. Eur J Pharmacol 500, 51-62.

Hayes, C. E., Nashold, F. E., *et al.*,2003. The immunological functions of the vitamin D endocrine system. Cell Mol Biol (Noisy-le-grand) 49, 277-300.

Haynie, S. L., Crum, G. A., *et al.*,1995. Antimicrobial activities of amphiphilic peptides covalently bonded to a water-insoluble resin. Antimicrob Agents Chemother 39, 301-307.

Heinz, S., Haehnel, V., *et al.*,2003. Species-specific regulation of Toll-like receptor 3 genes in men and mice. J Biol Chem 278, 21502-21509.

Hertz, G. Z. and Stormo, G. D.,1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15, 563-577.

Heuchel, R., Radtke, F.*, et al.*,1994. The transcription factor MTF-1 is essential for basal and heavy metal-induced metallothionein gene expression. Embo J 13, 2870-2875.

Hinton, B. T., Lan, Z. J.*, et al.*,1998. Testicular regulation of epididymal gene expression. J Reprod Fertil Suppl 53, 47-57.

Hiroi, M. and Ohmori, Y.,2003. The transcriptional coactivator CREB-binding protein cooperates with STAT1 and NF-kappa B for synergistic transcriptional activation of the CXC ligand 9/monokine induced by interferon-gamma gene. J Biol Chem 278, 651-660.

Hocker, M., Raychowdhury, R.*, et al.*,1998. Sp1 and CREB mediate gastrin-dependent regulation of chromogranin A promoter activity in gastric carcinoma cells. J Biol Chem 273, 34000-34007.

Holick, M. F.,2003. Evolution and function of vitamin D. Recent Results Cancer Res 164, 3-28.

Hongbiao, W., Baolong, N.*, et al.*,2005. Biological activities of cecropin B-thanatin hybrid peptides. J Pept Res 66, 382-386.

Horton, R. M.,1999. Scripting Wizards for Chime and RasMol. Biotechniques 26, 874-876.

Hostens, K., Pavlovic, D.*, et al.*,1999. Exposure of human islets to cytokines can result in disproportionately elevated proinsulin release. J Clin Invest 104, 67-72.

Huang, C. J., Nazarian, R., *et al.*,2002. Tumor necrosis factor modulates transcription of myelin basic protein gene through nuclear factor kappa B in a human oligodendroglioma cell line. Int J Dev Neurosci 20, 289-296.

Huang, H., Scherman, M. S., *et al.*,2005. Identification and Active Expression of the Mycobacterium tuberculosis Gene Encoding 5-Phospho-{alpha}-D-ribose-1-diphosphate: Decaprenyl-phosphate 5-Phosphoribosyltransferase, the First Enzyme Committed to Decaprenylphosphoryl-D-arabinose Synthesis. J Biol Chem 280, 24539-24543.

Hughes, A. L.,1999. Evolutionary diversification of the mammalian defensins. Cell Mol Life Sci 56, 94-103.

Ibrahim, H. R., Inazaki, D., *et al.*,2005. Processing of lysozyme at distinct loops by pepsin: a novel action for generating multiple antimicrobial peptide motifs in the newborn stomach. Biochim Biophys Acta 1726, 102-114.

Ihmels, J., Bergmann, S., *et al.*,2004. Defining transcription modules using large-scale gene expression data. Bioinformatics 20, 1993-2003.

Iida, K. and Nishimura, I.,2002. Gene expression profiling by DNA microarray technology. Crit Rev Oral Biol Med 13, 35-50.

Iida, K. T., Shimano, H., *et al.*,2001. Insulin up-regulates tumor necrosis factor-alpha production in macrophages through an extracellular-regulated kinase-dependent pathway. J Biol Chem 276, 32531-32537.

Imler, J. L. and Bulet, P.,2005. Antimicrobial peptides in Drosophila: structures, activities and gene regulation. Chem Immunol Allergy 86, 1-21.

Jang, B. C., Lim, K. J., *et al.*,2004. Up-regulation of human beta-defensin 2 by interleukin-1beta in A549 cells: involvement of PI3K, PKC, p38 MAPK, JNK, and NF-kappaB. Biochem Biophys Res Commun 320, 1026-1033.

Jeay, S., Sonenshein, G. E., *et al.*,2002. Growth hormone can act as a cytokine controlling survival and proliferation of immune cells: new insights into signaling pathways. Mol Cell Endocrinol 188, 1-7.

Jenab, S. and Inturrisi, C. E.,1995. Proenkephalin gene expression: interaction of glucocorticoid and cAMP regulatory elements. Biochem Biophys Res Commun 210, 589-599.

Jia, J., Yang, L., *et al.*,2006. EHPred: an SVM-based method for epoxide hydrolases recognition and classification. J Zhejiang Univ Sci B 7, 1-6.

Johansson, O., Alkema, W., *et al.*,2003. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. Bioinformatics 19 Suppl 1, i169-176.

Joshi, J. and Sabol, S. L.,1991. Proenkephalin gene expression in C6 rat glioma cells: potentiation of cyclic adenosine 3',5'-monophosphate-dependent transcription by glucocorticoids. Mol Endocrinol 5, 1069-1080.

Juvvadi, P., Vunnam, S., *et al.*,1999. Structure-activity studies of normal and retro pig cecropin-melittin hybrids. J Pept Res 53, 244-251.

Kadonaga, J. T.,2004. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell 116, 247-257.

Kamysz, W.,2005. Are antimicrobial peptides an alternative for conventional antibiotics? Nucl Med Rev Cent East Eur 8, 78-86.

Kamysz, W., Okroj, M., *et al.*,2003. Novel properties of antimicrobial peptides. Acta Biochim Pol 50, 461-469.

Kao, C. Y., Chen, Y., *et al.*,2004. IL-17 markedly up-regulates beta-defensin-2 expression in human airway epithelium via JAK and NF-kappaB signaling pathways. J Immunol 173, 3482-3491.

Kapetanovic, I. M., Rosenfeld, S., *et al.*,2004. Overview of commonly used bioinformatics methods and their applications. Ann N Y Acad Sci 1020, 10-21.

Karasavvas, K. A., Baldock, R., *et al.*,2004. Bioinformatics integration and agent technology. J Biomed Inform 37, 205-219.

Karp, P. D., Paley, S., *et al.*,2001. Database verification studies of SWISS-PROT and GenBank. Bioinformatics 17, 526-532; discussion 533-524.

Khanna-Gupta, A., Zibello, T., *et al.*,2000. Sp1 and C/EBP are necessary to activate the lactoferrin gene promoter during myeloid differentiation. Blood 95, 3734-3741.

King, A. E., Morgan, K., *et al.*,2003. Differential regulation of secretory leukocyte protease inhibitor and elafin by progesterone. Biochem Biophys Res Commun 310, 594-599.

Kingsley-Kallesen, M. L., Kelly, D., *et al.*,1999. Transcriptional regulation of the transforming growth factor-beta2 promoter by cAMP-responsive element-binding protein (CREB) and activating transcription factor-1 (ATF-1) is modulated by protein kinases and the coactivators p300 and CREB-binding protein. J Biol Chem 274, 34020-34028.

Kinugasa, T., Sakaguchi, T., *et al.*,2000. Claudins regulate the intestinal barrier in response to immune mediators. Gastroenterology 118, 1001-1011.

Klamt, S., Saez-Rodriguez, J., *et al.*,2006. A methodology for the structural and functional analysis of signaling and regulatory networks. BMC Bioinformatics 7, 56.

Kloster, M., Tang, C., *et al.*,2005. Finding regulatory modules through large-scale gene-expression data analysis. Bioinformatics 21, 1172-1179.

Kobierski, L. A., Wong, A. E., *et al.*,1999. Cyclic AMP-dependent activation of the proenkephalin gene requires phosphorylation of CREB at serine-133 and a Src-related kinase. J Neurochem 73, 129-138.

Koczulla, A. R. and Bals, R.,2003. Antimicrobial peptides: current status and therapeutic potential. Drugs 63, 389-406.

Konradi, C., Macias, W., *et al.*,2003. Striatal proenkephalin gene induction: coordinated regulation by cyclic AMP and calcium pathways. Brain Res Mol Brain Res 115, 157-161.

Kragol, G., Lovas, S., *et al.*,2001. The antibacterial peptide pyrrhocoricin inhibits the ATPase actions of DnaK and prevents chaperone-assisted protein folding. Biochemistry 40, 3016-3026.

Krause, A., Sillard, R., *et al.*,2003. Isolation and biochemical characterization of LEAP-2, a novel blood peptide expressed in the liver. Protein Sci 12, 143-152.

Krijgsveld, J., Zaat, S. A., *et al.*,2000. Thrombocidins, microbicidal proteins from human blood platelets, are C-terminal deletion products of CXC chemokines. J Biol Chem 275, 20374-20381.

Kumar, S., Tamura, K., *et al.*,2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5, 150-163.

Kylsten, P., Samakovlis, C., *et al.*,1990. The cecropin locus in Drosophila; a compact gene cluster involved in the response to infection. Embo J 9, 217-224.

Lawrence, C. E., Altschul, S. F., *et al.*,1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262, 208-214.

Le, Y., Gagneten, S., *et al.*,2003. Far-upstream elements are dispensable for tissue-specific proenkephalin expression using a Cre-mediated knock-in strategy. J Neurochem 84, 689-697.

Lee, T. L., Alba, D., *et al.*,2006. Application of transcriptional and biological network analyses in mouse germ-cell transcriptomes. Genomics 88, 18-33.

Lemaitre, B., Reichhart, J. M., *et al.*,1997. Drosophila host defense: differential induction of antimicrobial peptide genes after infection by various classes of microorganisms. Proc Natl Acad Sci U S A 94, 14614-14619.

Lennartsson, A., Pieters, K., *et al.*,2003. AML-1, PU.1, and Sp3 regulate expression of human bactericidal/permeability-increasing protein. Biochem Biophys Res Commun 311, 853-863.

Li, W. F., Ma, G. X., *et al.*,2006. Apidaecin-type peptides: biodiversity, structure-function relationships and mode of action. Peptides 27, 2350-2359.

Li, X., Massa, P. E., *et al.*,2002. IKKalpha, IKKbeta, and NEMO/IKKgamma are each required for the NF-kappa B-mediated inflammatory response program. J Biol Chem 277, 45129-45140.

Liu, D., Cai, S., *et al.*,2003. C1 inhibitor prevents endotoxin shock via a direct interaction with lipopolysaccharide. J Immunol 171, 2594-2601.

Liu, F., Kondova, I., *et al.*,2000. Detection of PACH1, a nuclear factor implicated in the transcriptional regulation of meiotic and early haploid stages of spermatogenesis. Mol Reprod Dev 57, 224-231.

Liu, K., Catalfamo, M., *et al.*,2002. IL-15 mimics T cell receptor crosslinking in the induction of cellular proliferation, gene expression, and cytotoxicity in CD8+ memory T cells. Proc Natl Acad Sci U S A 99, 6192-6197.

Lu, Z., Kim, K. A., *et al.*,2004. MEF up-regulates human beta-defensin 2 expression in epithelial cells. FEBS Lett 561, 117-121.

Luders, T., Birkemo, G. A., *et al.*,2003. Strong synergy between a eukaryotic antimicrobial peptide and bacteriocins from lactic acid bacteria. Appl Environ Microbiol 69, 1797-1799.

Ludwig, A., Berkhout, T., *et al.*,2002. Fractalkine is expressed by smooth muscle cells in response to IFN-gamma and TNF-alpha and is modulated by metalloproteinase activity. J Immunol 168, 604-612.

Ma, H., Ke, Y., *et al.*,2000. Bovine and human insulin activate CD8+-autoreactive CTL expressing both type 1 and type 2 cytokines in C57BL/6 mice. J Immunol 164, 86-92.

Ma, Y., Su, Q., *et al.*,1998. Differentiation-stimulated activity binds an ETS-like, essential regulatory element in the human promyelocytic defensin-1 promoter. J Biol Chem 273, 8727-8740.

Macian, F., Lopez-Rodriguez, C., *et al.*,2001. Partners in transcription: NFAT and AP-1. Oncogene 20, 2476-2489.

Maget-Dana, R.,1999. The monolayer technique: a potent tool for studying the interfacial properties of antimicrobial and membrane-lytic peptides and their interactions with lipid membranes. Biochim Biophys Acta 1462, 109-140.

Maggio, E. T. and Ramnarayan, K.,2001. Recent developments in computational proteomics. Trends Biotechnol 19, 266-272.

Maglott, D., Ostell, J., *et al.*,2005. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 33, D54-58.

Mahapatra, N. R., Mahata, M., *et al.*,2003. Secretin activation of chromogranin A gene transcription. Identification of the signaling pathways in cis and in trans. J Biol Chem 278, 19986-19994.

Mahata, S. K., Mahapatra, N. R., *et al.*,2002. Neuroendocrine cell type-specific and inducible expression of chromogranin/secretogranin genes: crucial promoter motifs. Ann N Y Acad Sci 971, 27-38.

Mallow, E. B., Harris, A., *et al.*,1996. Human enteric defensins. Gene structure and developmental expression. J Biol Chem 271, 4038-4045.

Matsuno, H., Doi, A., *et al.*,2000. Hybrid Petri net representation of gene regulatory network. Pac Symp Biocomput 341-352.

Matsushima-Nishiu, M., Unoki, M., *et al.*,2001. Growth and gene expression profile analyses of endometrial cancer cells expressing exogenous PTEN. Cancer Res 61, 3741-3749.

Matsuzaki, K.,1999. Why and how are peptide-lipid interactions utilized for self-defense? Magainins and tachyplesins as archetypes. Biochim Biophys Acta 1462, 1-10.

Maxwell, A. I., Morrison, G. M., *et al.*,2003. Rapid sequence divergence in mammalian beta-defensins by adaptive evolution. Mol Immunol 40, 413-421.

McKenzie, G., Ward, G., *et al.*,2006. Cellular Notch responsiveness is defined by phosphoinositide 3-kinase-dependent signals. BMC Cell Biol 7, 10.

Michel, T., Reichhart, J. M., *et al.*,2001. Drosophila Toll is activated by Gram-positive bacteria through a circulating peptidoglycan recognition protein. Nature 414, 756-759.

Mohamed-Ali, V., Pinkney, J. H., *et al.*,1998. Adipose tissue as an endocrine and paracrine organ. Int J Obes Relat Metab Disord 22, 1145-1158.

Moon, S. K., Lee, H. Y., *et al.*,2002. Activation of a Src-dependent Raf-MEK1/2-ERK signaling pathway is required for IL-1alpha-induced upregulation of beta-defensin 2 in human middle ear epithelial cells. Biochim Biophys Acta 1590, 41-51.

Moran, L. B., Duke, D. C., *et al.*,2004. Towards a transcriptome definition of microglial cells. Neurogenetics 5, 95-108.

Morris, M. A. and Ley, K.,2004. Trafficking of natural killer cells. Curr Mol Med 4, 431-438.

Morrison, G. M., Semple, C. A., *et al.*,2003. Signal sequence conservation and mature peptide divergence within subgroups of the murine beta-defensin gene family. Mol Biol Evol 20, 460-470.

Moss AC, D. P., MacMathuna P,2007. In Silico Promoter Analysis can Predict Genes of Functional Relevance in Cell Proliferation: Validation in a Colon Cancer Model. Translational oncogenomics 1-6.

Muczynski, K. A., Anderson, S. K.*, et al.*,1998. Discoordinate surface expression of IFN-gamma-induced HLA class II proteins in nonprofessional antigen-presenting cells with absence of DM and class II colocalization. J Immunol 160, 3207-3216.

Muller, C. A., Markovic-Lipkovski, J.*, et al.*,2002. Human alpha-defensins HNPs-1, -2, and -3 in renal cell carcinoma: influences on tumor cell proliferation. Am J Pathol 160, 1311-1324.

Murphy, B. J., Andrews, G. K.*, et al.*,1999. Activation of metallothionein gene expression by hypoxia involves metal response elements and metal transcription factor-1. Cancer Res 59, 1315-1322.

Murphy, C. J., Foster, B. A.*, et al.*,1993. Defensins are mitogenic for epithelial cells and fibroblasts. J Cell Physiol 155, 408-413.

Musikacharoen, T., Matsuguchi, T.*, et al.*,2001. NF-kappa B and STAT5 play important roles in the regulation of mouse Toll-like receptor 2 gene expression. J Immunol 166, 4516-4524.

Naitza, S. and Ligoxygakis, P.,2004. Antimicrobial defences in Drosophila: the story so far. Mol Immunol 40, 887-896.

Nason, G. P.,1996. Wavelet Shrinkage Using Cross-Validation. Journal of the Royal Statistical Society Series B (Methodological) 58, 463-479.

Nguyen, H., Teskey, L.*, et al.*,1999. Identification of the secretory leukocyte protease inhibitor (SLPI) as a target of IRF-1 regulation. Oncogene 18, 5455-5463.

Nicolas, G., Bennoun, M., *et al.*,2001. Lack of hepcidin gene expression and severe tissue iron overload in upstream stimulatory factor 2 (USF2) knockout mice. Proc Natl Acad Sci U S A 98, 8780-8785.

Nishikawa, J., Kitaura, M., *et al.*,1995. Vitamin D receptor contains multiple dimerization interfaces that are functionally different. Nucleic Acids Res 23, 606-611.

Niyonsaba, F., Hirata, M., *et al.*,2003. Epithelial cell-derived antibacterial peptides human beta-defensins and cathelicidin: multifunctional activities on mast cells. Curr Drug Targets Inflamm Allergy 2, 224-231.

Noble, W. S., Kuehn, S., *et al.*,2005. Predicting the in vivo signature of human gene regulatory sequences. Bioinformatics 21 Suppl 1, i338-343.

Nykanen, A., Vesanen, S., *et al.*,1998. Synergistic antimicrobial effect of nisin whey permeate and lactic acid on microbes isolated from fish. Lett Appl Microbiol 27, 345-348.

Odom, A., Muir, S., *et al.*,1997. Calcineurin is required for virulence of Cryptococcus neoformans. Embo J 16, 2576-2589.

Okumura, S., Kashiwakura, J., *et al.*,2003. Identification of specific gene expression profiles in human mast cells mediated by Toll-like receptor 4 and FcepsilonRI. Blood 102, 2547-2554.

Olsen, N. J. and Kovacs, W. J.,2001. Effects of androgens on T and B lymphocyte development. Immunol Res 23, 281-288.

O'Neil, D. A.,2003. Regulation of expression of beta-defensins: endogenous enteric peptide antibiotics. Mol Immunol 40, 445-450.

O'Neil, D. A., Porter, E. M., *et al.*,1999. Expression and regulation of the human beta-defensins hBD-1 and hBD-2 in intestinal epithelium. J Immunol 163, 6718-6724.

Oswald, F., Dobner, T., *et al.*,1996. The E2F transcription factor activates a replication-dependent human H2A gene in early S phase of the cell cycle. Mol Cell Biol 16, 1889-1895.

Ouellette, A. J.,1997. Paneth cells and innate immunity in the crypt microenvironment. Gastroenterology 113, 1779-1784.

Ovadia, H., Magenheim, Y., *et al.*,1996. Molecular characterization of immune derived proenkephalin mRNA and the involvement of the adrenergic system in its expression in rat lymphoid cells. J Neuroimmunol 68, 77-83.

Paone, G., Wada, A., *et al.*,2002. ADP ribosylation of human neutrophil peptide-1 regulates its biological properties. Proc Natl Acad Sci U S A 99, 8231-8235.

Park, C. B., Kim, M. S., *et al.*,1996. A novel antimicrobial peptide from Bufo bufo gargarizans. Biochem Biophys Res Commun 218, 408-413.

Pastorino, F., Brignole, C., *et al.*,2004. Targeted delivery of oncogene-selective antisense oligonucleotides in neuroectodermal tumors: therapeutic implications. Ann N Y Acad Sci 1028, 90-103.

Patil, A., Hughes, A. L., *et al.*,2004. Rapid evolution and diversification of mammalian alpha-defensins as revealed by comparative analysis of rodent and primate genes. Physiol Genomics 20, 1-11.

Pazgier, M., Hoover, D. M., *et al.*,2006. Human beta-defensins. Cell Mol Life Sci 63, 1294-1313.

Pereira, H. A.,2006. Novel therapies based on cationic antimicrobial peptides. Curr Pharm Biotechnol 7, 229-234.

Persad, S., Troussard, A. A.*, et al.*,2001. Tumor suppressor PTEN inhibits nuclear accumulation of beta-catenin and T cell/lymphoid enhancer factor 1-mediated transcriptional activation. J Cell Biol 153, 1161-1174.

Persson, P., Manetopoulos, C.*, et al.*,2004. Olf/EBF proteins are expressed in neuroblastoma cells: potential regulators of the Chromogranin A and SCG10 promoters. Int J Cancer 110, 22-30.

Petit, P., Bertrand, G.*, et al.*,1989. Effects of extracellular adenine nucleotides on the electrical, ionic and secretory events in mouse pancreatic beta-cells. Br J Pharmacol 98, 875-882.

Pohl, T. M., Phillips, E.*, et al.*,1990. The organisation of the mouse chromogranin B (secretogranin I) gene. FEBS Lett 262, 219-224.

Quinlan, K. L., Naik, S. M.*, et al.*,1999. Substance P activates coincident NF-AT- and NF-kappa B-dependent adhesion molecule gene expression in microvascular endothelial cells through intracellular calcium mobilization. J Immunol 163, 5656-5665.

Ramsay, R. G.,2005. c-Myb a stem-progenitor cell regulator in multiple tissue compartments. Growth Factors 23, 253-261.

Ravasi, T., Hsu, K.*, et al.*,2004. Probing the S100 protein family through genomic and functional analysis. Genomics 84, 10-22.

Reichardt, H. M.,2004. Immunomodulatory activities of glucocorticoids: insights from transgenesis and gene targeting. Curr Pharm Des 10, 2797-2805.

Reichardt, H. M., Tronche, F., *et al.*,2000. Molecular genetic analysis of glucocorticoid signaling using the Cre/loxP system. Biol Chem 381, 961-964.

Ren, B., Robert, F., *et al.*,2000. Genome-wide location and function of DNA binding proteins. Science 290, 2306-2309.

Rice, W. G., Ganz, T., *et al.*,1987. Defensin-rich dense granules of human neutrophils. Blood 70, 757-765.

Risso, A.,2000. Leukocyte antimicrobial peptides: multifunctional effector molecules of innate immunity. J Leukoc Biol 68, 785-792.

Rivas-Santiago, B., Schwander, S. K., *et al.*,2005. Human {beta}-defensin 2 is expressed and associated with Mycobacterium tuberculosis during infection of human alveolar epithelial cells. Infect Immun 73, 4505-4511.

Rodriguez de la Vega, R. C. and Possani, L. D.,2005. On the evolution of invertebrate defensins. Trends Genet 21, 330-332.

Roth, F. P., Hughes, J. D., *et al.*,1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 16, 939-945.

Rozansky, D. J., Wu, H., *et al.*,1994. Glucocorticoid activation of chromogranin A gene expression. Identification and characterization of a novel glucocorticoid response element. J Clin Invest 94, 2357-2368.

Sakaue, S., Nishihira, J., *et al.*,1999. Regulation of macrophage migration inhibitory factor (MIF) expression by glucose and insulin in adipocytes in vitro. Mol Med 5, 361-371.

Salzet, M.,2001. Neuroimmunology of opioids from invertebrates to human. Neuro Endocrinol Lett 22, 467-474.

Sandberg, M. B., Bloksgaard, M., *et al.*,2005. The gene encoding acyl-CoA-binding protein is subject to metabolic regulation by both sterol regulatory element-binding protein and peroxisome proliferator-activated receptor alpha in hepatocytes. J Biol Chem 280, 5258-5266.

Saugar, J. M., Rodriguez-Hernandez, M. J., *et al.*,2006. Activity of cecropin A-melittin hybrid peptides against colistin-resistant clinical strains of Acinetobacter baumannii: molecular basis for the differential mechanisms of action. Antimicrob Agents Chemother 50, 1251-1256.

Scheetz, T., Bartlett, J. A., *et al.*,2002. Genomics-based approaches to gene discovery in innate immunity. Immunol Rev 190, 137-145.

Schlezinger, J. J., Jensen, B. A., *et al.*,2002. Peroxisome proliferator-activated receptor gamma-mediated NF-kappa B activation and apoptosis in pre-B cells. J Immunol 169, 6831-6841.

Schmid, M., Fellermann, K., *et al.*,2004. The role of defensins in the pathogenesis of chronic-inflammatory bowel disease. Z Gastroenterol 42, 333-338.

Schorey, J. S. and Cooper, A. M.,2003. Macrophage signalling upon mycobacterial infection: the MAP kinases lead the way. Cell Microbiol 5, 133-142.

Schutte, B. C., Mitros, J. P., *et al.*,2002. Discovery of five conserved beta -defensin gene clusters using a computational search strategy. Proc Natl Acad Sci U S A 99, 2129-2133.

Schwartz, D. R., Wu, R., *et al.*,2003. Novel candidate targets of beta-catenin/T-cell factor signaling identified by gene expression profiling of ovarian endometrioid adenocarcinomas. Cancer Res 63, 2913-2922.

Schwarz, K., Eggers, M., *et al.*,2000. The proteasome regulator PA28alpha/beta can enhance antigen presentation without affecting 20S proteasome subunit composition. Eur J Immunol 30, 3672-3679.

Segal, E., Yelensky, R., *et al.*,2003. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics 19 Suppl 1, i273-282.

Selsted, M. E. and Ouellette, A. J.,2005. Mammalian defensins in the antimicrobial immune response. Nat Immunol 6, 551-557.

Selsted, M. E., Tang, Y. Q., *et al.*,1993. Purification, primary structures, and antibacterial activities of beta-defensins, a new family of antimicrobial peptides from bovine neutrophils. J Biol Chem 268, 6641-6648.

Semple, C. A., Rolfe, M., *et al.*,2003. Duplication and selection in the evolution of primate beta-defensin genes. Genome Biol 4, R31.

Shai, Y.,1999. Mechanism of the binding, insertion and destabilization of phospholipid bilayer membranes by alpha-helical antimicrobial and cell non-selective membrane-lytic peptides. Biochim Biophys Acta 1462, 55-70.

Shai, Y.,2002. Mode of action of membrane active antimicrobial peptides. Biopolymers 66, 236-248.

Sherman, H., Chapnik, N., *et al.*,2006. Albumin and amino acids upregulate the expression of human beta-defensin 1. Mol Immunol 43, 1617-1623.

Shike, H., Shimizu, C., *et al.*,2004. Organization and expression analysis of the zebrafish hepcidin gene, an antimicrobial peptide gene conserved among vertebrates. Dev Comp Immunol 28, 747-754.

Shinnar, A. E., Butler, K. L., *et al.*,2003. Cathelicidin family of antimicrobial peptides: proteolytic processing and protease resistance. Bioorg Chem 31, 425-436.

Sica, A., Dorman, L., *et al.*,1997. Interaction of NF-kappaB and NFAT with the interferon-gamma promoter. J Biol Chem 272, 30412-30420.

Siggia, E. D.,2005. Computational methods for transcriptional regulation. Curr Opin Genet Dev 15, 214-221.

Simmaco, M., Mignogna, G., *et al.*,1998. Antimicrobial peptides from amphibian skin: what do they tell us? Biopolymers 47, 435-450.

Singh, P. K., Jia, H. P., *et al.*,1998. Production of beta-defensins by human airway epithelia. Proc Natl Acad Sci U S A 95, 14961-14966.

Sinha, S., van Nimwegen, E., *et al.*,2003. A probabilistic method to detect regulatory modules. Bioinformatics 19 Suppl 1, i292-301.

Skerlavaj, B., Romeo, D., *et al.*,1990. Rapid membrane permeabilization and inhibition of vital functions of gram-negative bacteria by bactenecins. Infect Immun 58, 3724-3730.

Slutsky, S. G., Kamaraju, A. K., *et al.*,2003. Activation of myelin genes during transdifferentiation from melanoma to glial cell phenotype. J Biol Chem 278, 8960-8968.

Sonnenburg, S., Zien, A., *et al.*,2006. ARTS: accurate recognition of transcription starts in human. Bioinformatics 22, e472-480.

Staal, F. J. and Clevers, H. C.,2005. WNT signalling and haematopoiesis: a WNT-WNT situation. Nat Rev Immunol 5, 21-30.

Su, K., Li, X., *et al.*,2004. A promoter haplotype of the immunoreceptor tyrosine-based inhibitory motif-bearing FcgammaRIIb alters receptor expression and associates with autoimmunity. II. Differential binding of GATA4 and Yin-Yang1 transcription factors and correlated receptor expression and function. J Immunol 172, 7192-7199.

Suico, M. A., Koga, T., *et al.*,2004. Sp1 is involved in the transcriptional activation of lysozyme in epithelial cells. Biochem Biophys Res Commun 324, 1302-1308.

Taguchi, Y. and Imai, H.,2006. Expression of beta-defensin-2 in human gingival epithelial cells in response to challenge with Porphyromonas gingivalis in vitro. J Periodontal Res 41, 334-339.

Takeuchi, A., Reddy, G. S., *et al.*,1998. Nuclear factor of activated T cells (NFAT) as a molecular target for 1alpha,25-dihydroxyvitamin D3-mediated effects. J Immunol 160, 209-218.

Teng, C. T.,2002. Lactoferrin gene expression and regulation: an overview. Biochem Cell Biol 80, 7-16.

Tennessen, J. A.,2005. Molecular evolution of animal antimicrobial peptides: widespread moderate positive selection. J Evol Biol 18, 1387-1394.

Thevissen, K., Cammue, B. P., *et al.*,2000. A gene encoding a sphingolipid biosynthesis enzyme determines the sensitivity of Saccharomyces cerevisiae to an antifungal plant defensin from dahlia (Dahlia merckii). Proc Natl Acad Sci U S A 97, 9531-9536.

Thomas, M. D., Kremer, C. S., *et al.*,2005. c-Myb is critical for B cell development and maintenance of follicular B cells. Immunity 23, 275-286.

Thompson, J. D., Gibson, T. J., *et al.*,1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25, 4876-4882.

Thompson, J. D., Higgins, D. G., *et al.*,1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673-4680.

Tompa, M., Li, N., *et al.*,2005. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23, 137-144.

Tourkova, I. L., Shurin, G. V., *et al.*,2005. Restoration by IL-15 of MHC class I antigen-processing machinery in human dendritic cells inhibited by tumor-derived gangliosides. J Immunol 175, 3045-3052.

Trappe, R., Doenecke, D., *et al.*,1999. The expression of human H2A-H2B histone gene pairs is regulated by multiple sequence elements in their joint promoters. Biochim Biophys Acta 1446, 341-351.

Tsutsumi-Ishii, Y., Hasebe, T., *et al.*,2000. Role of CCAAT/enhancer-binding protein site in transcription of human neutrophil peptide-1 and -3 defensin genes. J Immunol 164, 3264-3273.

Uhl, G. R., Appleby, D., *et al.*,1991. Synaptic regulation of the enkephalin gene and transcription factors in vivo: possible roles in drug abuse. NIDA Res Monogr 105, 123-129.

v Agoston, D., Santha, E., *et al.*,1998. Isolation and structural and genetic analysis of the mouse enkephalin gene and its d(AC/TG)n repeats. DNA Seq 9, 217-226.

Valdar, W. S.,2002. Scoring residue conservation. Proteins 48, 227-241.

van den Berg, R. H., Faber-Krol, M. C., *et al.*,1998. Inhibition of activation of the classical pathway of complement by human neutrophil defensins. Blood 92, 3898-3903.

van Helden, J.,2003. Regulatory sequence analysis tools. Nucleic Acids Res 31, 3593-3596.

van 't Hof, W., Veerman, E. C., *et al.*,2001. Antimicrobial peptides: properties and applicability. Biol Chem 382, 597-619.

Veiga, D. F., Vicente, F. F., *et al.*,2006. Gene networks as a tool to understand transcriptional regulation. Genet Mol Res 5, 254-268.

Vizioli, J. and Salzet, M.,2002. Antimicrobial peptides from animals: focus on invertebrates. Trends Pharmacol Sci 23, 494-496.

von Horsten, H. H., Derr, P., *et al.*,2002. Novel antimicrobial peptide of human epididymal duct origin. Biol Reprod 67, 804-813.

Vora, P., Youdim, A., *et al.*,2004. Beta-defensin-2 expression is regulated by TLR signaling in intestinal epithelial cells. J Immunol 173, 5398-5405.

Voss, E., Wehkamp, J., *et al.*,2006. NOD2/CARD15 mediates induction of the antimicrobial peptide human beta-defensin-2. J Biol Chem 281, 2005-2011.

Wade, D., Andreu, D., *et al.*,1992. Antibacterial peptides designed as analogs or hybrids of cecropins and melittin. Int J Pept Protein Res 40, 429-436.

Wang, T. T., Nestel, F. P., *et al.*,2004. Cutting edge: 1,25-dihydroxyvitamin D3 is a direct inducer of antimicrobial peptide gene expression. J Immunol 173, 2909-2912.

Wang, W., Cherry, J. M., *et al.*,2005. Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. Proc Natl Acad Sci U S A 102, 1998-2003.

Wang, X., Zhang, Z., *et al.*,2003. Airway epithelia regulate expression of human beta-defensin 2 through Toll-like receptor 2. Faseb J 17, 1727-1729.

Wang, Y., Wimmer, U., *et al.*,2004. Metal-responsive transcription factor-1 (MTF-1) is essential for embryonic liver development and heavy metal detoxification in the adult liver. Faseb J 18, 1071-1079.

Wang, Z. and Wang, G.,2004. APD: the Antimicrobial Peptide Database. Nucleic Acids Res 32, D590-592.

Wasserman, W. W. and Fickett, J. W.,1998. Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol 278, 167-181.

Wasserman, W. W. and Sandelin, A.,2004. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5, 276-287.

Wehkamp, J., Harder, J., *et al.*,2004. NF-kappaB- and AP-1-mediated induction of human beta defensin-2 in intestinal epithelial cells by Escherichia coli Nissle 1917: a novel effect of a probiotic bacterium. Infect Immun 72, 5750-5758.

Wehkamp, J., Salzman, N. H., *et al.*,2005. Reduced Paneth cell alpha-defensins in ileal Crohn's disease. Proc Natl Acad Sci U S A 102, 18129-18134.

Wei, Q., Miskimins, W. K., *et al.*,2003. Cloning and characterization of the rat myelin basic protein gene promoter. Gene 313, 161-167.

Wei, Q., Miskimins, W. K., *et al.*,2004. Sox10 acts as a tissue-specific transcription factor enhancing activation of the myelin basic protein gene promoter by p27Kip1 and Sp1. J Neurosci Res 78, 796-802.

Wei, Q., Miskimins, W. K., *et al.*,2005. Stage-specific expression of myelin basic protein in oligodendrocytes involves Nkx2.2-mediated repression that is relieved by the Sp1 transcription factor. J Biol Chem 280, 16284-16294.

Welling, M. M., Paulusma-Annema, A., *et al.*,2000. Technetium-99m labelled antimicrobial peptides discriminate between bacterial infections and sterile inflammations. Eur J Nucl Med 27, 292-301.

Werhli, A. V., Grzegorczyk, M., *et al.*,2006. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. Bioinformatics 22, 2523-2531.

Werner, T.,1999. Models for prediction and recognition of eukaryotic promoters. Mamm Genome 10, 168-175.

Werner, T.,2001. Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data. Pharmacogenomics 2, 25-36.

Werner, T.,2002. Finding and decrypting of promoters contributes to the elucidation of gene function. In Silico Biol 2, 249-255.

Werner, T.,2003. Promoters can contribute to the elucidation of protein function. Trends Biotechnol 21, 9-13.

Werner, T., Fessele, S., *et al.*,2003. Computer modeling of promoter organization as a tool to study transcriptional coregulation. Faseb J 17, 1228-1237.

Wheeler, D. L., Barrett, T., *et al.*,2005. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 33, D39-45.

Williams, M. J.,2001. Regulation of antibacterial and antifungal innate immunity in fruitflies and humans. Adv Immunol 79, 225-259.

Wingender, E., Chen, X., *et al.*,2000. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 28, 316-319.

Witthoft, T., Pilz, C. S., *et al.*,2005. Enhanced human beta-defensin-2 (hBD-2) expression by corticosteroids is independent of NF-kappaB in colonic epithelial cells (CaCo2). Dig Dis Sci 50, 1252-1259.

Wu, H., Zhang, G., *et al.*,2000. Regulation of cathelicidin gene expression: induction by lipopolysaccharide, interleukin-6, retinoic acid, and Salmonella enterica serovar typhimurium infection. Infect Immun 68, 5552-5558.

Xiang, C. C. and Chen, Y.,2000. cDNA microarray technology and its applications. Biotechnol Adv 18, 35-46.

Xiao, Y., Hughes, A. L., *et al.*,2004. A genome-wide screen identifies a single beta-defensin gene cluster in the chicken: implications for the origin and evolution of mammalian defensins. BMC Genomics 5, 56.

Yamaguchi, Y., Nagase, T., *et al.*,2002. Identification of multiple novel epididymis-specific beta-defensin isoforms in humans and mice. J Immunol 169, 2516-2523.

Yamamoto, C. M., Banaiee, N., *et al.*,2004. Alpha-defensin expression during myelopoiesis: identification of cis and trans elements that regulate expression of NP-3 in rat promyelocytes. J Leukoc Biol 75, 332-341.

Yang, D., Chen, Q., *et al.*,2000. Human neutrophil defensins selectively chemoattract naive T and immature dendritic cells. J Leukoc Biol 68, 9-14.

Yang, L., Weiss, T. M., *et al.*,2000. Crystallization of antimicrobial pores in membranes: magainin and protegrin. Biophys J 79, 2002-2009.

Yasin, B., Pang, M., *et al.*,2000. Evaluation of the inactivation of infectious Herpes simplex virus by host-defense peptides. Eur J Clin Microbiol Infect Dis 19, 187-194.

Yudkin, J. S., Kumari, M., *et al.*,2000. Inflammation, obesity, stress and coronary heart disease: is interleukin-6 the link? Atherosclerosis 148, 209-214.

Zaballos, A., Villares, R., *et al.*,2004. Identification on mouse chromosome 8 of new beta-defensin genes with regionally specific expression in the male reproductive organ. J Biol Chem 279, 12421-12426.

Zasloff, M.,2002. Antimicrobial peptides of multicellular organisms. Nature 415, 389-395.

Zhang, B., Georgiev, O., *et al.*,2003. Activity of metal-responsive transcription factor 1 by toxic heavy metals and H2O2 in vitro is modulated by metallothionein. Mol Cell Biol 23, 8471-8485.

Zhang, M. Q.,2002. Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet 3, 698-709.

Zhang, Z. and Gerstein, M.,2003. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J Biol 2, 11.

Zhou, C. X., Zhang, Y. L.*, et al.*,2004. An epididymis-specific beta-defensin is important for the initiation of sperm maturation. Nat Cell Biol 6, 458-464.

Supplementary Material

We are all inventors, each sailing out on a voyage of discovery, guided each by a private chart, of which there is no duplicate. The world is all gates, all opportunities. (Ralph Waldo Emerson)

Supplementary Tables and Figures for Chapter 5

Supplementary Table 5.1 AMPcg families and representative members in mouse, rat and human

Mm: *Mus musculus*; Hs: *Homo sapiens*; Rn: *Rattus norvegicus*: TUID: transcriptional unit ID; CTSS: transcription start site (TSS) information based on CAGE tags.

AMP Family	Gene symbol	Species	Representative CloneID/Accession	TUID	CTSS
Alpha defensin	2010016B13Rik	Mm	2010016B13	175722	No
	2010016F14Rik	Mm	2010016F14	168136	No
	defa5	Hs	NM_021010	-	-
	defa6	Hs	NM_001926	-	-
	defa4	Hs	NM_001925	-	-
	defa3	Hs	NM_005217	-	-
Apoa2 (apolipoprotein A-II)	Apoa2	Mm	I530003A11	83109	No
	APOA2	Hs	HIT000032344.2	-	-
	Apoa2	Rn	NM_013112	-	-
Beta defensin	9230107O10Rik	Mm	9230107O10	103672	No
	DEFB28	Hs	AF525930	-	-
	Defb1	Mm	D630029A12	169116	Yes
	DEFB1	Hs	BC033298	-	-
	Defb1	Rn	NM_031810	-	-
	Defb23	Mm	1700012K18	121132	Yes
	DEFB123	Hs	NM_153324	-	-
	Defb4	Mm	2310001F05	168175	No
	DEFB4	Hs	NM_004942	-	-
	Defb36	Mm	1700011J22	168985	Yes
	DEFB105a	Hs	NM_152250	-	-
	Defb12	Mm	9230103N16	77756	No
	Defb19	Mm	4930563B01	81337	No

BPI (Bactericidal/permeability-increasing)	9230105K17Rik	Mm	9230105K17	112251	No
	BPI	Hs	BC040955	-	-
Bin1b/SPAG11	Spag11	Mm	9230111C08	168760	No
	SPAG11	Hs	NM_016512	-	-
	Spag11	Rn	NM_145087	-	-
Cathelicidin	Camp	Mm	F930015N03	112000	Yes
	CAMP	Hs	NM_004345	-	-
	cramp	Rn	AF484553	-	-
Calgranulin	S100a9	Mm	F430201H11	83114	Yes
	S100a9	Hs	NM_002965	-	-
	S100a9	Rn	NM_053587	-	-
DBI (Acyl-CoA-binding protein family)	Dbi	Mm	6720460E16	102356	Yes
	DBI	Hs	NM_020548	-	-
Slpi (skin-derived antileukoproteinase)	Slpi	Mm	2310075E18	75903	No
	SLPI	Hs	HIT000038907.2	-	-
	Slpi	Rn	NM_053372	-	-
Granulin	Grn	Mm	0610012H06	104193	Yes
	GRN	Hs	BC000324	-	-
	Grn	Rn	NM_017113	-	-
Hepcidin	1810073K19Rik	Mm	2210420P15	168118	Yes
	LEAP2	Hs	NM_052971	-	-
Histone 2A derived defense peptide	Hist1h2ac	Mm	9030420B16	112273	No
	HIST1H2AC	Hs	NM_003512	-	-
	Hist1h2ae	Mm	1190022L06	112736	No
	HIST1H2AE	Hs	NM_021052	-	-
Lactoferrin	Ltf	Mm	9830118D19	173811	No
	LTF	Hs	NM_002343	-	-

Lysozyme	9530003J23Rik	Mm	9530003J23	106239	No
	Lyzs	Mm	I420013M05	111075	Yes
	LYZS	Hs	AF099029	-	-
	Lyzs	Rn	NM_012771	-	-
MBP (Myelin Basic Protein)	Prg2	Mm	2510004C07	112877	No
	PRG2	Hs	HIX0009634.2	-	-
	prg2	Rn	NM_031619	-	-
Melanotropin alpha (Pro-opiomelanocortin family)	Pomc1	Mm	5730403F20	151196	No
	POMC1	Hs	NM_000939	-	-
	Pomc1	Rn	NM_139326	-	-
PENK (Proenkaphalin) (opioid neuropeptide family)	Penk1	Mm	4922504O09	179452	Yes
	PENK	Hs	HIX0007519.2	-	-
	Penk-rs	Rn	NM_017139	-	-
Secretogranin I (chromogranin/secretogranin family)	Chgb	Mm	5730420J08	177050	Yes
	CHGB	Hs	HIX0015625.2	-	-
	Chgb	Rn	NM_012526	-	-
SPYY (Skin peptide tyrosine-tyrosine) (NPY family)	Npy	Mm	0710005A05	72959	Yes
	Pyy	Mm	C820007C10	111251	Yes
	NPY	Hs	HIX0006525.2	-	-
	Npy	Rn	NM_012614	-	-

Vasostatin (Chromogranin A) (chromogranin/secretogranin family)	Chga	Mm	G630083O06	83089	Yes
	CHGA	Hs	HIX0011909.2	-	-
	Chga	Rn	NM_021655	-	-
VIP (Vasoactive intestinal peptide) (Glucagon family)	Vip	Mm	9130007F05	112113	No
	VIP	Hs	HIX0006306.2	-	-
ZAP (CCCH type, antiviral 1)	Zc3hav1	Mm	F420004O17	99218	Yes
	ZC3HAV1	Hs	HIX0007129.3	-	-
	Zap	Rn	NM_173045	-	-

Supplementary Table 5.2 FANTOM3 dataset-derived AMP transcripts which were new to mouse and absent in human

Riken clone ID/ GenBank acession	Gene Symbol
D730003B11	*1..1.1.1.1.1.1 Csnd*
D730017I01	*Csnd*
D730018F19	*Csnd*
D730018I02	*Csnd*
D730032K03	*Csnd*
D730045O16	*Csnd*
D730048M03	*Csnd*
2010300L12	*Defcr-rs1*
2010319H24	*Defcr-rs1*
5033416M10	*Mcpt2*
G630050E22	*Mcpt4*
9030622B11	*Mcpt8*
0610031H01	*Hist2h2aa2*
1700048I17	*Hist2h2aa2*

Supplementary Table 5.3 TFs associated with *ab initio*-predicted TFBSs that coincided with experimental data.

References are cited at the end of Supplementary material section

AMP family	Experimentally determined TFs	References	Predicted TFs matching experimental confirmed TFs
Alpha defensin	CAAT, PEBP2/CBF	(Yamamoto *et al.*, 2004)	PEBP2/CBF
Beta defensin	NF-KAPPAB, AP-1, NF-IL6, MEF, VDR	(Harder *et al.*, 2000),(Vora *et al.*, 2004),(Lu *et al.*, 2004), (Wang T.T. *et al.*, 2004)	AP-1,MEF(C-ETS1),VDR
BPI	AML-1, PU.1, SP3/SP1,C/EBP,USF, NF-KB, C-REL	(Lennartsson *et al.*, 2003)	SP3/SP1,AML-1,NF-KB
Cathelicidin	VDR, NF-IL6, RAR,IL-6RE	(Wang T.T. *et al.*, 2004), (Frohm Nilsson *et al.*, 1999), (Wu *et al.*, 2000)	VDR, RAR, IL-6 RE
DBI	SREBP, SP1, PPAR-ALPHA,AP-1, C/EBP, HNF-3, RXR-ALPHA, NF-1/CTF , AP-2	(Sandberg *et al.*, 2005), (Elholm *et al.*, 1996)	SREBP, SP1,AP-1,RXR-ALPHA,NF-1/CTF
Hepcidin	C/EBP-ALPHA	(Courselaud *et al.*, 2002)	C/EBPALPHA
Histone 2A	TBP,OCT-1,CAAT box,	(Oswald *et al.*, 1996), (Albig *et al.*, 1999), (Trappe *et al.*, 1999)	Oct-1, CAAT
Lactoferrin	SP1, C/EBP	(Teng, 2002), (Khanna-Gupta *et al.*, 2000)	SP1, C/EBP
Lysozyme	SP1,MEF,C/EBP	(Suico *et al.*, 2004)	MEF (C-ETS1),SP-1
MBP	NKX2.2,SP1,SOX10,PAX3, NF-KB	(Wei *et al.*, 2005), (Wei *et al.*, 2004), (Wei *et al.*, 2003), (Slutsky *et al.*, 2003), (Huang *et al.*, 2002)	SP1
Melanotropin-alpha	SRE,AP-1, AP-2 LIKE, CAAT BOX	(Deen *et al.*, 1991)	AP-1, AP-2 LIKE
Proenkaphalin1	TATA, AP-2, NF-KAPPAB,MZF-1,MYC PACH1,CREB,CRE, NF1,AP-1	(Liu *et al.*, 2000), (Kobierski *et al.*, 1999), (Fu *et al.*, 1997) (Le *et al.*, 2003)	AP-1, NF1, TATA,AP2,NF-KB,MZF-1,NF-Y,
Secretogranin I	SP-1, CRE, TATA	(Pohl *et al.*, 1990), (Mahata *et al.*, 2002)	SP-1, TATA
Vasostatin	OLF/EBF, SP1,CREB,GR	(Persson *et al.*, 2004), (Mahapatra *et al.*, 2003), (Hocker *et al.*, 1998), (Rozansky *et al.*, 1994)	GR,SP-1
VIP	OCT-1,MEF-2,STAT,AP-1,CRE	(Hahm and Eiden, 1998)	STAT1,AP-1,POUF1A(OCT-

			1)
Slpi	GR,PR, IRF-1	(Hayashi *et al.*, 2004), (King *et al.*, 2003), (Nguyen *et al.*, 1999)	GR, PR
Apoa2	NA	NA	NA
Calgranulin	NA	NA	NA
Granulin	NA	NA	NA
SPYY	NA	NA	NA
ZAP	NA	NA	NA
Bin1b/SPAG11	NA	NA	NA

Supplementary Table 5.4 Total number of motifs found for each AMP family

Unknown: motif does not match any of the TRANSFAC-listed TF binding sites

AMP family	new TFs	Unknown motifs	total
Alpha-defensin	73	3	77
Apoa2	36	6	42
BPI	113	4	120
Beta-defensin	78	8	89
Bin1b/SPAG11	75	3	78
Calgranulin	162	4	166
Cathelicidin	75	3	81
DBI	53	1	59
Granulin	67	3	70
Hepcidin	59	3	63
Histone 2A	83	12	97
Lactoferrin	46	4	52
Lysozyme	30	9	41
MBP	67	2	70
Melanotropin alpha	81	1	84
Proenkaphalin1	54	3	85
SPYY	58	5	63
Secretogranin I	31	6	39
Slpi	94	5	101
VIP	54	3	60
Vasostatin	19	9	30
ZAP	77	5	82

Supplementary Table 5.5. Ranking of TF groups according to their frequency of appearance in different AMP families.
For example, under rank 1, AD is the the most frequently occurring TF group in five of the AMP families that are listed in Table 5.5. Underlined numbers indicate the TF groups that are high-ranking such as liver-specific, nervous system-related, adipocyte-related, nuclear hormone-related, immune cell-specific and lung-specific TFs.

Rank	Tissue/Function-specific TF groups											
	AD	**NHR**	**CC**	**IMM**	**LIV**	**LUNG**	**MUS**	**NS**	**PAN**	**PIT**	**EYE**	**BS**
1	5	6	5	6	9	5	1	7	0	1	0	0
2	6	6	2	2	6	3	0	4	1	0	0	0
3	4	1	1	2	5	1	1	3	1	0	0	0
4	3	2	0	0	1	1	0	3	1	1	0	0
5	0	3	2	4	1	4	3	5	2	2	0	0
6	1	0	1	5	0	3	4	0	1	2	0	0
7	0	1	4	3	0	2	4	0	5	2	0	0
8	2	1	3	0	0	2	1	0	5	4	0	0
9	0	1	4	0	0	1	4	0	2	3	0	2
10	1	1	0	0	0	0	4	0	4	6	0	4
11	0	0	0	0	0	0	0	0	0	1	14	16
12	0	0	0	0	0	0	0	0	0	0	8	0
Avg.	3.32	3.55	5.27	3.95	2.05	4.32	7.05	2.77	7.18	7.77	11.36	10.64

Supplementary Table 5.6: Ranksum test of AMPcg families versus house keeping genes

Unadjusted p-value is p-value from the ranksum test for each AMP family vs. house keeping gene set. Adjusted p-value: p-value corrected for multiplicity testing for 6 TF groups that are being tested: namely, AD, NHR, IMM, LIV, LUNG, NS, these are the most frequently present TF groups across all AMP families. Seven AMP families have corrected P-value less than 0.05 and thus the null hypothesis that the two distributions are the same can be rejected. This means the distirbution of these families of TFs and those in the testes AMP families are different.

AMP Families	AD	NHR	CC	IMM	LIV	LUNG	MUS	NS	PAN	PIT	EYE	BS	Unadjusted p-value	Adjusted p-value
Alphadefensin	6	12	2	6	6	5	3	9	2	1	0	0	0.5591	1
Apoa2	5	5	5	4	5	5	4	6	4	3	0	0	0.0074	0.0444
Betadefensin	6	5	4	6	6	5	2	8	2	3	0	1	0.124	0.744
bin1b/spag11	9	9	3	5	10	10	6	10	5	2	0	3	0.7614	1
Bpi	6	7	8	8	8	5	8	7	7	4	0	0	0.3293	1
Calgranulin	8	5	6	9	9	6	7	11	7	5	1	2	0.8165	1
Cathelicidin	4	8	6	8	5	5	1	6	3	5	0	1	0.1503	0.9018
Dbi	7	8	4	5	7	6	4	6	1	4	0	0	0.2576	1
Slpi	6	7	3	6	6	5	3	8	5	4	0	0	0.2199	1
Granulin	6	5	6	4	6	6	4	5	4	3	0	0	0.0269	0.1614
Hepcidin	10	9	3	7	11	11	3	9	6	7	0	0	0.4387	1
Histone	5	2	3	3	5	4	3	3	4	3	0	0	0.0004	0.0024
Lactoferrin	7	10	3	4	7	6	4	8	3	1	0	0	0.3922	1
Lysozyme	4	2	4	3	4	4	3	3	2	1	0	0	0	0
Mbp	6	7	9	9	7	6	6	7	2	6	0	2	0.3562	1
Melanotropinalpha	9	6	9	7	8	7	4	8	3	4	0	0	0.5525	1
Proenkaphalin	7	7	3	4	7	7	1	8	3	3	0	1	0.2883	1
Secretogranin	1	5	2	6	3	2	4	3	3	3	0	1	0.0013	0.0078
Spyy	5	5	2	5	5	3	1	5	3	5	0	0	0.002	0.012
Vip	3	3	3	4	3	2	1	3	3	1	0	1	0	0
Vstn	4	4	3	3	5	4	3	4	1	2	0	0	0.0001	0.0006
Zap	7	8	6	6	8	8	3	6	4	4	0	0	0.4358	1
Control set														
House_Keeping_genes	4	12	5	14	7	6	4	7	4	5	0	1		

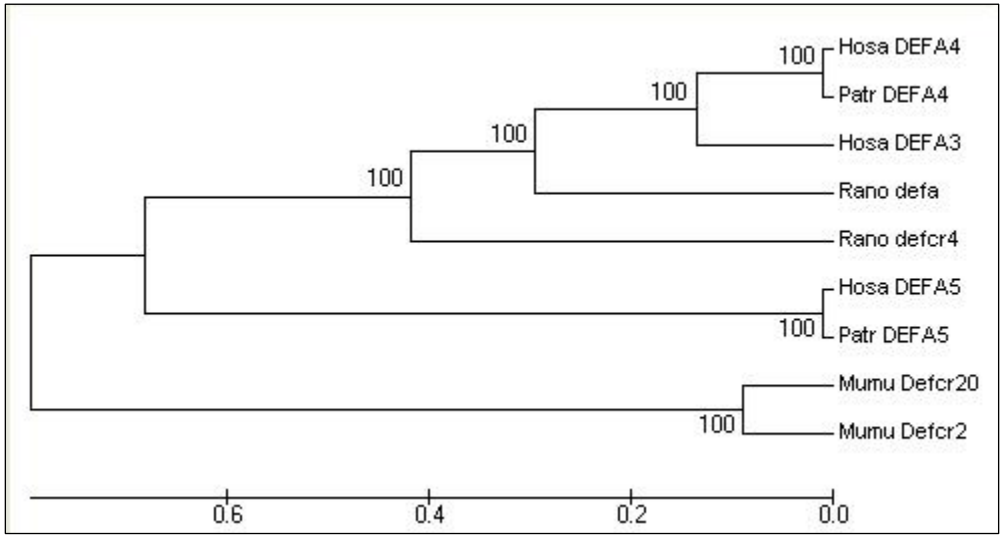
Supplementary Table 5.7 P-value table of motif groups.

The row with bold-face values indicates the boundary of eleven AMP families that were significantly enriched in predicted NHR-binding motifs relative to the whole AMP family set.

No. of NHR binding motif candidates in subpopulation	No. of motifs from all families included in subpopulations	No. of NHRbinding motif candiates	Total population (motifs from all families)	Bonferroni correction factor	P-value	Bonferroni corrected p-value	No. of groups included in statistically significant set
137	420	139	440	440	0.023154572	1	21
135	400	139	440	440	0.000884858	0.389337708	20
132	380	139	440	440	0.000128334	0.056466838	19
128	360	139	440	440	5.62042E-05	0.024729858	18
123	340	139	440	440	5.99005E-05	0.026356217	17
118	320	139	440	440	4.74915E-05	0.020896267	16
113	300	139	440	440	3.08239E-05	0.013562527	15
108	280	139	440	440	1.69494E-05	0.00745775	14
103	260	139	440	440	7.94636E-06	0.0034964	13
98	240	139	440	440	3.14129E-06	0.001382167	12
92	**220**	**139**	**440**	**440**	**2.81167E-06**	**0.001237134**	**11**
85	200	139	440	440	5.55134E-06	0.002442591	10
78	180	139	440	440	9.01694E-06	0.003967454	9
71	160	139	440	440	1.22648E-05	0.005396502	8
64	140	139	440	440	1.39428E-05	0.006134828	7
56	120	139	440	440	3.3908E-05	0.014919519	6
48	100	139	440	440	6.8438E-05	0.030112705	5
40	80	139	440	440	0.000112308	0.049415705	4
31	60	139	440	440	0.000399135	0.175619353	3
22	40	139	440	440	0.001106333	0.486786677	2
12	20	139	440	440	0.006893247	1	1

Figures for Chapter 5

Supplementary Figure 5.1. UPGMA tree for alpha-defensin promoter regions analyzed in this study
The tree topology coincides with the exception of rat Defcr4 with the previously reported enteric (i.e. intestine) and myeloid/neutrophil cell expression of rat, mouse and human alpha-defensins. The cluster comprising *Hosa-defa3, -defa4, Patr-defa4* and Rano-Defa represents myeloid-specific alpha-defensins. *Mumu-Defcr20, -Defcr2, Rano-Defcr4, Hosa-defa5* and *Patr-defa5* represent the enteric-expressed group of alpha-defensins. The species abbreviations are Mumu: *Mus musculus*; Hosa: *Homo sapiens*; Patr: *Pan troglodytes*; Rano: *Rattus norvegicus*.



Supplementary Tables and Figures for Chapter 6

Supplementary Table 6.1 TFs that correspond to *ab-initio* predicted motifs derived from Penk family promoter regions.

All motifs were detected in mouse, rat and human sequences. The underlined TF binding sites are known to bind TFs in the proenkephalin promoter region (Liu *et al.*, 2000, Kobierski *et al.*, 1999, Fu *et al.*, 1997, Le *et al.*, 2003). The species abbreviations are Hs: *Homo sapiens*; Mm: *Mus musculus*: Rn: *Rattus norvegicus*. Unknown: motif does not match any of the TRANSFAC-listed TF binding sites.

Motif no.	Motif occurance	Species	Motif	TF name
1	3	Mm,Hs,Rn	CCAGTAACCTGCG	FXR:RXR-alpha LXR-alpha:RXR-alpha LXR-beta:RXR-alpha ERRalpha1
2	3	Mm,Hs,Rn	TATAAAGTGGCTGT	TFIID TBP
3	3	Mm,Hs,Rn	GATCTAAAGAAGAAA	AR GR
4	3	Mm,Hs,Rn	CCAAGTCCGTC	SF-1 GR
5	3	Mm,Hs,Rn	TTAAGATCCCCA	NF-kappaB1 NF-kappaB2 NF-kappaB2 precursor AP-2alpha AP-2alphaA
6	3	Mm,Hs,Rn	GTGATDCAGGA	AP-1 c-Fos c-Jun JunD
7	3	Mm,Hs,Rn	TCCAGVAAGDH	c-Ets-1 Elk-1 SAP-1a SAP-1b SRF PEA3 ELF-1
8	3	Mm,Hs,Rn	CAGGCGTCGGCGCG	DREB1A ZF5 E2F
9	3	Mm,Hs,Rn	CGATTGGGGCGCGC	NFI/CTF CTF NF-Y
10	3	Mm,Hs,Rn	CCAGAVAGGCAG	UBP-1 GATA-1 GATA-3 Meis-1a Meis-1b GATA-4 RXR-beta VDR MOT3
11	3	Mm,Hs,Rn	CCGGGTCCTA	Unknown
12	3	Mm,Hs,Rn	AGCCCGTGBC	USF-1 USF1 USF2 USF2b USF HMBP EmBP-1a
13	3	Mm,Hs,Rn	GTGACTTTGCCCCA	DSF GCN4 COUP-TF1 RAR-beta RXR-alpha RAR-alpha1 TLX Pax-2.1 Pax-2.2 IRF-4 IRF-8 AP-2alpha AP-2alphaA C/EBPgamma PPAR-gamma:RXR-alpha VDR LXR-alpha:RXR-alpha
14	3	Mm,Hs,Rn	GATCTGTBTT	Sox2 Meis-1a Meis-1b GR
15	3	Mm,Hs,Rn	TGAAATTTGG	Unknown
16	3	Mm,Hs,Rn	GCTGTGGGGACGTCC	AML1 AML1a AML1c MZF1 MIG1 MZF-1 AP-2alpha AP-2alphaA MBP-1 (1) NF-kappaB1 NF-kappaB2 NF-kappaB2 precursor
17	3	Mm,Hs,Rn	BHHCAAGAGGA	Unknown
18	3	Mm,Hs,Rn	GGAAGGGGCAG	VDR LXR-alpha:RXR-alpha CAC-binding protein NF-E2 PPAR-gamma:RXR-alpha Sp1
19	3	Mm,Hs,Rn	AHGCCCCAACC	Sp1 PPAR-gamma:RXR-alpha VDR LXR-alpha:RXR-alpha AP-2alphaA ADR1 C/EBPalpha C/EBPbeta
20	3	Mm,Hs,Rn	GGACAGGATG	Meis-1a Meis-1b Elk-1 SAP-1a SAP-1b SRF E47 Fli-1 Net TCF

Supplementary Table 6.2 TF binding sites that correspond to *ab-initio*-predicted motifs derived from Zap family promoter regions.

The species abbreviations are Hs: *Homo sapiens,* Mm: *Mus musculus*, Rn: *Rattus norvegicus*. Unknown: motif does not match any of the TRANSFAC-listed TF binding sites.

.Motif No	Motif occurrence	Species	Motif	TF binding sites
1	3	Mm,Hs,Rn	CTCCACCTGTTCCTT	Alfin1,RXR-alpha,VDR,E12,E47,MyoD,myogenin,EMF1,EMF2,EMF3,EMF4,Myf-5,c-Myc,USF2,CAN,E2A,DEP2,HEB ,Ac,AS-C T3,Da,Sc,Sn, CLIM2,GATA-1,Lmo2,Tal-1,USF-1,NeuroD,NEUROD,LVa,PR B,AR,GR,c-Ets-2,ESE-1,HELIOS,LyF-1
2	3	Mm,Hs, Rn	TCACCGCACT	ER-alpha,ABI4,AML1a
3	3	Mm,Hs, Rn	CTGGGGGGCCC	MIG1,Sp1,ZAC-1a
4	3	Mm,Hs, Rn	AAGCAGTTGGT	c-Myb,c-Myc,E47,NeuroD,NEUROD,E12,MyoD,MyoD:E12,myogenin,Myogenin:E12,Dec-02,c-Myb:HES-1
5	3	Mm,Hs, Rn	GGCTCTTTAATT	AR,GR,LF-A1,RAR-alpha1,RAR-beta,RAR-gamma,RORalpha1,RXR-beta2,LXR-alpha:RXR-alpha,LXR-beta:RXR-alpha, T3R-alpha,FXR:RXR-alpha, PXR-1:RXR-alpha,COUP,FOR1,FOR2,ER-alpha,AP-1,RXR-alpha,TAF(II)28,LXR-alpha,VDR,TR2-11,PPAR-gamma,
6	3	Mm,Hs, Rn	CATGACCCTGGAG	RXR-gamma,CAR:RXR-alpha,Nkx2-1
7	3	Mm,Hs, Rn	ACTCTAAGGTAT	Unknown
8	3	Mm,Hs, Rn	ATTCGCTCTCCC	LyF-1,RXR-beta,VDR
9	3	Mm,Hs, Rn	GGTTTACCTT	CAR:RXR-alpha,LXR-alpha:RXR-alpha,SXR,RAR-beta,RAR-gamma,RXR-alpha,RAR-alpha1,ER-alpha
10	3	Mm,Hs, Rn	GAGCGGCACC	Unknown
11	3	Mm,Hs, Rn	AATATCCAAG	NF-1,TGGCA-binding protein
12	3	Mm,Hs, Rn	AGCAGCATCA	Unknown
13	3	Mm,Hs, Rn	GAGAGTAACAA	GATA-6,GCN4,PR B
14	3	Mm,Hs, Rn	AATAGGACTT	GR
15	3	Mm,Hs, Rn	CGGATTTGAGGACGC	Unknown
16	3	Mm,Hs, Rn	AAAATCATCTT	Otx2,GATA-3
17	3	Mm, Rn	TAAGTTTCGATTCT	Unknown
18	3	Mm,Hs, Rn	GGAGTCTGGAGG	Nkx2-1
19	3	Mm, Rn	GAGTTGGAAAGCGA	NF-AT1,NF-1,Ftz
20	3	Mm,Hs, Rn	GTGCGCCCACGG	MTF-1

Supplementary Tables and Figures for Chapter 7

Supplementary Table 7.1: Specificity and Sensitivity of the promoter models

AMP group	Gene names	Refseq Id	Species	Sensitivity	Specificity
alpha defensin 1	DEFA1	NM_004084	Hs	100	
	DEFA3	NM_005217	Hs	100	
	MNP1A	NM_001032862	Mmu	100	5/5+6
				Average(100%)	45.4%
alpha defensin 5	DEFA5	NM_021010	Hs	100	
	Defcr2	U03028	Mm	100	
	Defcr3	NM_007850	Mm	100	3/3+0
				Average (100%)	100%

The specificity was calculated with the formula: (Sp) = TP/ (TP+ FP) ; TP: True positive, FP: False Positive

Supplementary Table 7.2: Statistical significance of predicted genes from promoter model scan

The null hypothesis tested is that proportions A=k/n and B=K/N of genes are the same. The Bonferroni corrected p-value indicates that these two proportions are very different and implies that predictions of genes that are co-expressed by DEFA5 based on promoter model is very good (A >> B).

AMP model		
DEFA1	**Parameters**	**Values**
Total no. of promoters scanned	N	10255
Total no. genes predicted	K	104
Total no. of coexpressed genes that are present in the promoter dataset	n	51
Total no. of genes that matched coexpressed genes	k	17
	bonferroni correction factor	10255
		P_value = 3.62347894569854e-022, corrected P_value = 3.71587765881385e-018
DEFA5		
Total no. of promoters scanned	N	10255
Total no. genes predicted	K	240
Total no. of coexpressed genes that are present in the promoter dataset	n	226
Total no. of genes that matched coexpressed genes	k	177
	bonferroni correction factor	10255
		P_value = 1.07817800604295e-278, corrected P_value = 1.10567154519705e-274

Supplementary Table 7.3a: DEFA5 predicted genes that matched co-expression data

H-inv Ids: Unique gene identifier from H-Invitational database

H-inv IDs	Gene Names	Gene Description	Tissue	Pathway
HIT000036029	SNAI1	SimilartoEscargot/snail protein homolog(Fragment),partial cds.	Testis, embryonal carcinoma	Adherens junction
HIT000036885	SLUG	SimilartoSlugprotein,complete cds.	Uterus, leiomyosarcoma	Adherens junction
HIT000039321	TAF11	TBP-associatedfactor11;TAF11 RNApolymeraseII,TATAbox	Lung, small cell carcinoma	Basal transcription factors
HIT000031115	HER3	SimilartoReceptorprotein-tyrosinekinaseerbB-3 precursor(EC	Placenta, choriocarcinoma	Calcium signaling pathway
HIT000032887	CCNB	G2/mitotic-specificcyclinB1,partial cds.	Placenta, choriocarcinoma	Calcium signaling pathway,Cell cycle
HIT000036690	CLDN2	PMP-22/EMP/MP20andclaudinfamilyprotein, complete cds.	Colon, adenocarcinoma	Cell adhesion molecules (CAMs),Tight junction
HIT000036966	PIGT	Phosphatidylinositolglycan class T precursor (Homosapiens)	Skin, melanotic melanoma.	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
HIT000032824	PFKL	Phosphofructokinase, liver;	Lung, large cell carcinoma	Insulin signaling pathway,Galactose metabolism,Fructose and mannose metabolism,Glycolysis / Gluconeogenesis,Pentose phosphate pathway
HIT000031158	MNK1	Similar to MAPkinase-interactingserine/ threoninekinase1	Placenta, choriocarcinoma	Insulin signaling pathway,MAPK signaling pathway
HIT000002234	MAP2K1IP1	Mitogen-activated protein kinase kinase1i nteracting protein1;	bone marrow	MAPK signaling pathway
HIT000011725	WBSCR17	Similar to Williams-Beurensyndromecritical region gene17,	Brain	O-Glycan biosynthesis
HIT000036299	GFAP	Glialfibrillary acidic protein(Homosapiens),completecds.	Brain, glioblastoma with EGFR amplification	Prion disease,Neurodegenerative Disorders

HIT000042159	KIAA0106,PRDX6	Peroxiredoxin 6;antioxidantprotein2;non-seleniumglutathione	Brain	Prostaglandin and leukotriene metabolism,Alkaloid biosynthesis II,Methane metabolism,Phenylalanine metabolism,Stilbene, coumarine and lignin biosynthesis,2,4-Dichlorobenzoate degradation,Butanoate metabolism
HIT000036745	SPC18	Microsomal signalpeptidase 18kDasubunit	Colon, adenocarcinoma	Protein export
HIT000038280	TXNRD1	Similar to thioredoxinreductase1; KM-102-derivedreductase-like	Lung, large cell carcinoma	Pyrimidine metabolism
HIT000032334	PRIM1	DNA primases mallsubunit	Bone marrow, chronic myelogenous leukemia	Pyrimidine metabolism,Purine metabolism,DNA polymerase
HIT000031313	UGP2	SimilartoUTP--glucose-1-phosphateuridylyl transferase2	Lymph, Burkitt lymphoma	Starch and sucrose metabolism,Galactose metabolism,Nucleotide sugars metabolism,Pentose and glucuronate interconversions
HIT000029290	NUDT5	NudixhydrolaseNUDT5,partial cds.	Placenta, choriocarcinoma	Starch and sucrose metabolism,Purine metabolism,Folate biosynthesis
HIT000036950	MID1	Similar to DNA-3-methyladenine glycosylase (EC3.2.2.21)	Pancreas, epithelioid carcinoma	Tryptophan metabolism
HIT000038917	AUH	AU-bindingprotein/enoyl-CoAhydratase,complete cds.	Testis, embryonal carcinoma	Valine, leucine and isoleucine degradation
HIT000031341	NMP238	RuvB-like1 (EC3.6.1.-) (49-kDa TATAbox-binding)	Lung, small cell carcinoma	Wnt signaling pathway
HIT000001945	B5R1	Cytochromeb5reductase1,partial cds.	adrenal gland	
HIT000001890	HSPA14	Similar to Circadian OSCILLATORREGULATOR	adrenal gland	
HIT000002063	BC040106	Conservedhypotheticalprotein,completecds.	Blood	
HIT000042336	KIAA0198,PLAGL2	Similar toPLAGL2 (Pleiomorphicadenomagene-like2),partialcds.	bone marrow	
HIT000040748	H3F3A	HistoneH3familyprotein,complete cds.	Bone marrow, acute myelogenous leukemia	
HIT000037755	MRS1	Pannexin1,complete cds.	Bone, osteosarcoma	
HIT000000234	KIAA0517,NARF	Similar to Tripartitemotifprotein2 (Neuralactivity-related)	Brain	
HIT000000363	KIAA0638,	Pleckstrin-likedomaincontainingprotein,complete cds.	Brain	
HIT000001102	KIAA1387,	EVH1domaincontainingprotein,complete cds.	Brain	
HIT000001339	KIAA1624,	Conservedhypotheticalprotein,partial cds.	Brain	

HIT000001461	KIAA1745,KIAA1745	Semaphorin4Bprecursor,complete cds.	Brain	
HIT000025239	DKFZp564L023,UBQLN1	ubiquilin1 isoform1(Homosapiens),completecds.	Brain	
HIT000042245	KIAA0091,KIAA0091	Membrane-boundtranscriptionfactorsite-1protease precursor	Brain	
HIT000042312	KIAA0175,PK38	Similar to Proteinkinase PK38(Maternal embryonic leucinezipper)	Brain	
HIT000011833	LOC148137	Conserved hypothetical protein,partialcds.	Brain	
HIT000011902	BX647638	Zn-finger,C2H2 type domain containing protein,partial cds.	Brain	
HIT000012048	AF454939	Conserved hypothetica lprotein,complete cds.	Brain	
HIT000015894	FLJ33708	Hypothetical protein,complete cds.	Brain	
HIT000021463	KBTBD6	BTB/POZ domaincontainingprotein,partial cds.	Brain	
HIT000021509	RCBTB1	Regulatorofchromosomecondensation,RCC1 family protein,	Brain	
HIT000040515	RDH1	11-cis retinoldehydrogenase(EC1.1.1.105)(11-cisRDH),complete	Brain, anaplastic oligodendroglioma with 1p/19q loss	
HIT000035181	MGC17330	Kringledomaincontainingprotein,complete cds.	Brain, anaplastic oligodendroglioma with 1p/19q loss	
HIT000041611	GALT4	Beta-1,3-galactosyltransferase4 (EC2.4.1.62)(Beta-1,3-GalTase	Brain, fetal, whole pooled	
HIT000037657	TMSB10	thymosin, beta10(Homosapiens),completecds.	Brain, glioblastoma	
HIT000039547	LAG1	Similar to Longevity assurance homolog1(UOG-1protein)(LAG1)	Brain, hypothalamus	
HIT000041220	CDH14	Cadherin-18 precursor (Cadherin-14),complete cds.	Brain, hypothalamus	
HIT000039784	MKI67IP	Nucleolar protein interacting with the FHAdomain of pKi-67	Brain, hypothalamus	
HIT000030877	BGN	Biglycan precursor (Bone/cartilageproteoglycanI)(PG-S1),	Brain, neuroblastoma	
HIT000031004	CLG4A	72kDa typeIV collagenase precursor(EC3.4.24.24)	Brain, neuroblastoma	
HIT000033725	SPIN1	General substrate ransporter family protein,partial cds.	Brain, neuroblastoma	
HIT000033832	ATPAF1	ATP synthasemitochondrialF1complexassemblyfactor1;homolog	Brain, primitive neuroectodermal	
HIT000038173	RPP40	ribonucleaseP1;ribonucleaseP(40kD);ribonucleaseP,40kD	Brain, primitive neuroectodermal	
HIT000031528	TRX1	Thioredoxin(ATL-derivedfactor)(ADF)(Surfaceassociated	Cervix, carcinoma	
HIT000033663	BK215D111	RNA-bindingproteinregulatorysubunit,completecds.	Cervix, carcinoma	
HIT000034876	GT197	Beclin1(Coiled-coilmyosin-likeBCL2-interactingprotein)	Cervix, carcinoma	
HIT000038327	UCC1	Mammalianependyminrelatedprotein-1precursor(MERP-1)(UCC1	Cervix, carcinoma	
HIT000035576	FLJ20605	MOSCN-terminalbetabarreldomaincontainingprotein,complete	Cervix, carcinoma	
HIT000040926	7h3	RhoGAPdomaincontainingprotein,completecds.	Cervix, carcinoma	
HIT000007962	PTD015	PTD015protein(Homosapiens),completecds.	Colon	

HIT000008584	FLJ21657	Conservedhypotheticalprotein,completecds.	Colon	
HIT000032907	HOX1G	HomeoboxproteinHox-A9(Hox-1G),partialcds.	Colon, adenocarcinoma	
HIT000037949	APOL1	apolipoproteinL1isoformaprecursor;apolipoproteinL;	Colon, adenocarcinoma	
HIT000039048	SLC17A5	solutecarrierfamily17(anion/sugartransporter),member5;	Colon, adenocarcinoma	
HIT000041727	FABPL	Fattyacid-bindingprotein,liver(L-FABP),partialcds.	Colon, Kidney, Stomach, adult, whole pooled	
HIT000020679	C9orf150	Conservedhypotheticalprotein,completecds.	Heart	
HIT000027595	AL832683	Hypotheticalprotein,completecds.	human adipose	
HIT000027097	AL832185	Questionabletranscript,completesequence.	human cervix	
HIT000028510	DTX1	SimilartoDELTEX1,completecds.	human endometrium carcinoma cell line	
HIT000028526	SPATS2	Conservedhypotheticalprotein,completecds.	human endometrium carcinoma cell line	
HIT000009540	AK026266	C2domaincontainingprotein,partialcds.	human small intestine	
HIT000010243	HRPT2	RNApolIIaccessoryfactor,Cdc73familyprotein,partialcds.	human small intestine	
HIT000010276	DDX31	SimilartoRNAhelicase(Fragment),partialcds.	human small intestine	
HIT000002335	ARP11	Actin-relatedprotein10(hARP11),completecds.	Hypothalamus	
HIT000002429	CR612307	Conservedhypotheticalprotein,completecds.	Hypothalamus	
HIT000003106	PAK1IP1	PAK1interactingprotein1;PAK1-interactingprotein;	ileal mucosa	
HIT000002946	AK000471	Hypotheticalprotein,completecds.	ileal mucosa	
HIT000012430	PODXL	podocalyxin-likeprecursor;podocalyxin(Homosapiens),partial	Kidney	
HIT000016815	LTB4DH	SimilartoNADP-dependentleukotrieneB412-hydroxydehydrogenase	Kidney	
HIT000025305	DKFZp566N2024,NESH	NESHprotein;newmoleculeincludingSH3(Homosapiens),partial	Kidney	
HIT000035749	MAP17	17kDamembraneassociatedprotein(DD96protein),completecds.	Kidney, hypernephroma	
HIT000037427	MRPS36	Conservedhypotheticalprotein,completecds.	Kidney, hypernephroma	
HIT000032241	ISOT	Ubiquitincarboxyl-terminalhydrolase5(EC3.1.2.15)(Ubiquitin	Kidney, renal cell adenocarcinoma	
HIT000038939	HST	Alcoholsulfotransferase(EC2.8.2.2)(Hydroxysteroid	Liver	
HIT000020824	HSS	N-sulphoglucosaminesulphohydrolaseprecursor(EC3.10.1.1)	Lung	
HIT000020830	TYRO10	Discoidindomainreceptor2precursor(EC2.7.1.112)(Receptor	Lung	
HIT000034497	ATPIF1	MitochondrialATPaseinhibitor,IATPfamilyprotein,completecds.	Lung, large cell carcinoma	
HIT000040234	TRIM38	Zn-finger,RINGdomaincontainingprotein,completecds.	Lung, large cell carcinoma	
HIT000037644	HSPC117	ProteinofunknownfunctionUPF0027familyprotein,completecds.	Lung, mucoepidermoid carcinoma	
HIT000029637	PRDX2	peroxiredoxin2isoforma;thioredoxin-dependentperoxide	Lung, small cell carcinoma	

HIT000029675	DRG2	DevelopmentallyregulatedGTP-bindingprotein2(DRG2),partial	Lung, small cell carcinoma	
HIT000030859	HO2	Hemeoxygenase2(EC1.14.99.3)(HO-2),completecds.	Lung, small cell carcinoma	
HIT000032455	NEC2	Neuroendocrineconvertase2precursor(EC3.4.21.94)(NEC2)(PC2)	Lung, small cell carcinoma	
HIT000034802	TALDOR	Transaldolase(EC2.2.1.2),partialcds.	Lung, small cell carcinoma	
HIT000029571	ILF2	NF45protein,completecds.	Lung, small cell carcinoma	
HIT000031956	BC051849	Conservedhypotheticalprotein,completecds.	Lung, small cell carcinoma	
HIT000033592	MRPS2	RibosomalproteinS2,bacterialandorganelleformfamilyprotein,	Lung, small cell carcinoma	
HIT000037835	FLJ20013	2OG-Fe(II)oxygenasesuperfamilyprotein,completecds.	Lung, small cell carcinoma	
HIT000039948	NAALAD1	GlutamatecarboxypeptidaseII(EC3.4.17.21)(Membraneglutamate	Lung, Spleen, fetal, pooled	
HIT000039962	GRID	GRB2-relatedadaptorprotein2(GADSprotein)(Growthfactor	Lung, Spleen, fetal, pooled	
HIT000027727	TMEM30A	EukaryoticproteinofunknownfunctionDUF284familyprotein,	lymph node	
HIT000033379	SAP114	Splicingfactor3subunit1(Spliceosomeassociatedprotein114)	Lymph, Burkitt lymphoma	
HIT000036615	RF1	Eukaryoticpeptidechainreleasefactorsubunit1(eRF1)	Lymph, Burkitt lymphoma	
HIT000031846	AY736034	Cyclin-likeF-boxdomaincontainingprotein,completecds.	Lymph, Burkitt lymphoma	
HIT000040146	SATT	NeutralaminoacidtransporterA(SATT)(Alanine/serine/cysteine/	Lymph, lymphoma	
HIT000032665	AP2B1	adaptor-relatedproteincomplex2,beta1subunit;adaptin,beta2	Muscle, rhabdomyosarcoma	
HIT000033210	TOMM34	MitochondrialimportreceptorsubunitTOM34(Translocaseofouter	Muscle, rhabdomyosarcoma	
HIT000034687	STX6	Syntaxin6,completecds.	Muscle, rhabdomyosarcoma	
HIT000031771	NAT9	GCN5-relatedN-acetyltransferasedomaincontainingprotein,	Muscle, rhabdomyosarcoma	
HIT000003861	TMEM33	ProteinofunknownfunctionUPF0121familyprotein,completecds.	ovarian cancer	
HIT000003917	RNMTL1	tRNA/rRNAmethyltransferase(SpoU)familyprotein,completecds.	ovarian cancer	
HIT000030149	NIFIE14	SimilartoSeventransmembranedomainprotein,completecds.	Ovary, adenocarcinoma	
HIT000031192	CKLFSF6	chemokine-likefactorsuperfamily6(Homosapiens),completecds.	Ovary, adenocarcinoma	
HIT000034621	MRPL4	mitochondrialribosomalproteinL4isoforma(Homosapiens),	Ovary, adenocarcinoma	
HIT000032704	BRMS1L	Conservedhypotheticalprotein,completecds.	Ovary, adenocarcinoma	
HIT000006074	SCAND2	SimilartoSCANdomain-containingprotein2isoform2;SCAN	ovary, tumor tissue	
HIT000010815	MSTP028		ovary, tumor tissue	
HIT000039978	OSR1	odd-skippedrelated1;odz(oddOz/ten-m)related1(Homo	Pancreas, Spleen, adult pooled	
HIT000040025	G10P1	SimilartoInterferon-inducedproteinwithtetratricopeptide	Pancreas, Spleen, adult pooled	
HIT000004408	2410046H15RIK	SimilartoVitaminDreceptor-interactingproteincomplex	Placenta	
HIT000004626	STAU2	SimilartoDouble-strandedRNA-bindingproteinStaufen2long	Placenta	
HIT000030398	HN1	HN1protein(Hematologicalandneurologicalexpressed1protein),	Placenta, choriocarcinoma	

HIT000030438	COX4AL	NeighborofCOX4,completecds.	Placenta, choriocarcinoma	
HIT000032034	DIA1	NADH-cytochromeb5reductase(EC1.6.2.2)(B5R),partialcds.	Placenta, choriocarcinoma	
HIT000037948	HRBL	HIV-1Revbindingprotein-like;Rev/Rexactivationdomainbinding	Placenta, choriocarcinoma	
HIT000038651	TIN2	TERF1-interactingnuclearfactor2(TRF1-interactingnuclear	Placenta, choriocarcinoma	
HIT000038816	CTRP6	Complement-c1qtumornecrosisfactor-relatedprotein6precursor,	Placenta, choriocarcinoma	
HIT000030311	MGC5509	Conservedhypotheticalprotein,completecds.	Placenta, choriocarcinoma	
HIT000036138	C2orf30	Conservedhypotheticalprotein,partialcds.	Placenta, choriocarcinoma	
HIT000036609	TTF	Rho-relatedGTP-bindingproteinRhoH(GTP-bindingproteinTTF),	Primary B-Cells from Tonsils	
HIT000038424	DHLAG	HLAclassIIhistocompatibilityantigen,gammachain(HLA-DR	Primary B-Cells from Tonsils	
HIT000033818	BAP29	SimilartoB-cellreceptor-associatedprotein29(BCR-associated	Prostate	
HIT000034389	AIF1	SimilartoAllograftinflammatoryfactor-1(AIF-1)(Daintain),	Prostate	
HIT000035173	F3	Tissuefactorprecursor(TF)(CoagulationfactorIII)	Prostate, adenocarcinoma.	
HIT000041853	MBD1	methyl-CpGbindingdomainprotein1isoform3(Homosapiens),	Prostate, carcinoma	
HIT000017568	SAP61	SimilartoSplicingfactor3Asubunit3(Spliceosomeassociated	skeletal muscle	
HIT000033787	C20orf114	Lipid-bindingserumglycoproteinfamilyprotein,completecds.	Skeletal Muscle	
HIT000038179	MRPL46	Conservedhypotheticalprotein,completecds.	Skeletal Muscle	
HIT000035898	MOV34L	26Sproteasomenon-ATPaseregulatorysubunit7(26Sproteasome	Skin, melanotic melanoma, high MDR.	
HIT000039106	IFI30	interferon,gamma-inducibleprotein30preproprotein;	Skin, melanotic melanoma, high MDR.	
HIT000031594	TGT	Ubiquitincarboxyl-terminalhydrolase14(EC3.1.2.15)(Ubiquitin	Skin, melanotic melanoma.	
HIT000031679	LGALS4	Galectin-4(Lactose-bindinglectin4)(L-36lactosebinding	Skin, melanotic melanoma.	
HIT000032173	TSSC3	tumorsuppressingsubtransferablecandidate3;imprintedin	Skin, melanotic melanoma.	
HIT000034611	ARFL3	ADP-ribosylationfactor-likeprotein3,partialcds.	Skin, melanotic melanoma.	
HIT000034822	D17WSU104E	SimilartoDNAsegment,Chr17,WaynestateUniversity104,	Skin, melanotic melanoma.	
HIT000035389	COL9A3	Collagenalpha3(IX)chainprecursor,partialcds.	Skin, melanotic melanoma.	
HIT000036085	RPMS13	28SribosomalproteinS26,mitochondrialprecursor(MRP-S26)(MRP-	Skin, melanotic melanoma.	
HIT000031571	TBC1D17	RabGAP/TBCdomaincontainingprotein,completecds.	Skin, melanotic melanoma.	
HIT000021780	AK096925	Questionabletranscript.	small intestine	
HIT000021855	MTMR11	Conservedhypotheticalprotein,partialcds.	small intestine	
HIT000015332	FLJ00386,CTG7A	Positivecofactor2glutamine/Q-rich-associatedprotein(PC2	Spleen	
HIT000028634	MSZF13	SimilartoMszf13(Fragment),completecds.	Stomach	
HIT000014799	IOPPP	Inorganicpyrophosphatase(EC3.6.1.1)(Pyrophosphatephospho-	Testis	

HIT000018125	SOC	socius(Homosapiens),completecds.	Testis	
HIT000025493	DKFZp434E248,	GTP-bindingprotein,HSR1-relatedfamilyprotein,completecds.	Testis	
HIT000026715	DKFZp434C2120,KIAA1667	Hermansky-Pudlaksyndrome4protein(Light-earproteinhomolog),	Testis	
HIT000038574	LABH2	Abhydrolasedomaincontainingprotein2(ProteinPHPS1-2),	Testis	
HIT000040591	BC093018	Conservedhypotheticalprotein,partialcds.	Testis	
HIT000035842	HDHD1A	Haloaciddehalogenase-likehydrolasefamilyprotein,partialcds.	Testis, embryonal carcinoma	
HIT000038990	CCDC12	Conservedhypotheticalprotein,completecds.	Testis, embryonal carcinoma	
HIT000013018	FLJ31842	TRAM,LAG1andCLN8homologydomaincontainingprotein,complete	Tongue	
HIT000002015	LOC51255	SimilartoGenomicDNA,chromosome3,P1clone MSJ3,	umbilical cord blood	
HIT000031071	CEV14	SimilartoThyroidreceptorinteractingprotein11(TRIP-11)	Uterus, endometrium adenocarcinoma	
HIT000031086	DDX49	DEAD/DEAHboxhelicasedomaincontainingprotein,partialcds.	Uterus, endometrium adenocarcinoma	
HIT000031778	RNUT1	SNURPORTIN1(RNA,Utransporter1),completecds.	Uterus, endometrium adenocarcinoma	
HIT000035511	LOC113444	Conservedhypotheticalprotein,completecds.	Uterus, endometrium adenocarcinoma	
HIT000034247	GM2A	GangliosideGM2activatorprecursor(GM2-AP)(Cerebrosidesulfate	Uterus, leiomyosarcoma	
HIT000034834	SCAM1	SimilartoVinexin(SH3-containingadaptormolecule-1)(SCAM-1),	Uterus, leiomyosarcoma	
HIT000037450	TCF3G	Hepatocytenuclearfactor3-gamma(HNF-3G)(Forkheadboxprotein	Uterus, leiomyosarcoma	
HIT000038271	PIM2	Serine/threonine-proteinkinasePim-2(EC2.7.1.37)(Pim-2h),	Uterus, leiomyosarcoma	
HIT000033497	DENR	Density-regulatedproteinDRP1familyprotein,completecds.	Uterus, leiomyosarcoma	
HIT000035667	CBX6	Chromodomaincontainingprotein,completecds.	Uterus, leiomyosarcoma	
HIT000010634	D-UBP-64E	SimilartoUbiquitincarboxyl-terminalhydrolase64E(EC3.1.2.15)	whole embryo, mainly body	
HIT000007420	TMEM35	Conservedhypotheticalprotein,completecds.	whole embryo, mainly body	
HIT000003470	DERP7	Dermalpapilladerivedprotein7,completecds.	whole embryo, mainly head	
HIT000004995	COPS7A	SimilartoCOP9complexsubunit7A(COP9(Constitutive	whole embryo, mainly head	

Supplementary Table 7.3b: DEFA5 predicted genes that did not match co-expression data

H-inv IDs	Gene Names	Gene Description	Tissue	Pathway
HIT000002506	FAM45A	Conservedhypotheticalprotein,completecds.	adipose tissue	
HIT000001878	C9orf32	EukaryoticproteinofunknownfunctionDUF858familyprotein,	adrenal gland	
HIT000036637	MRPL20	RibosomalproteinL20familyprotein,completecds.	Bone marrow, acute myelogenous leukemia	
HIT000033794	TM4SF1	transmembrane4superfamilymember1;membranecomponent,	Bone marrow, chronic myelogenous leukemia	
HIT000020388	EFCBP1	EFhandcalciumbindingprotein1;synaptotagmininteracting	Brain	
HIT000042208	KIAA0049,M17S2	membranecomponent,chromosome17,surfacemarker2;1A1-3B;	Brain	
HIT000021327	TENC1	Hypotheticalprotein,completecds.	Brain	
HIT000040002	FLJ10560	Cyclase-associatedproteindomaincontainingprotein,completecds.	Brain, adult, 6 pooled whole brains	
HIT000035185	LRRN6A	Cysteine-richflankingregion,N-terminaldomaincontaining	Brain, anaplastic oligodendroglioma with 1p/19q loss	
HIT000037490	CX3CL1	chemokine(C-X3-Cmotif)ligand1;smallinduciblecytokine	Brain, neuroblastoma	Cytokine-cytokine receptor interaction
HIT000036260	TRIM9	Zn-finger,RINGdomaincontainingprotein,completecds.	Brain, neuroblastoma	
HIT000038709	CG13951/CG8803	SimilartoLethal(2)k10201protein(WunenregionBprotein),	Cervix, carcinoma	
HIT000007959	SCD5	SimilartoAcyl-CoA-desaturase,partialcds.	Colon	
HIT000040664	PDZK2	natrium-phosphatecotransporterIIaC-terminal-associatedprotein	Colon, Kidney, Stomach, adult, whole pooled	
HIT000019739	RASSF4	SimilartoRasandRabinteractor2(Rasinteraction/interference	corpus callosum	
HIT000021028		Conservedhypotheticalprotein,completecds.	Esophagus	
HIT000008909	SPBC15D416	SimilartoCellcyclecontrolproteincwf22,partialcds.	human small intestine	
HIT000002750	KIAA1125	ProteinkinaseCbindingprotein1(Rack7)(CutaneousT-cell	ileal mucosa	
HIT000002981	ZMYND13	SimilartoAnkyrinrepeatandMYNDdomaincontainingprotein1	ileal mucosa	

HIT000002897		Non-protein-codingtranscript,completesequence.	ileal mucosa	
HIT000002952		Hypotheticalprotein,completecds.	ileal mucosa	
HIT000032347	EPM2A	Similartoepilepsy,progressivemyoclonustype2A,Laforadisease	Kidney, hypernephroma	
HIT000029867	NANH	Sialidase1precursor(EC3.2.1.18)(Lysosomalsialidase)	Kidney, renal cell adenocarcinoma	
HIT000036261	H2AFX	H2AhistonefamilymemberX;H2AXhistone(Homosapiens),	Lung, small cell carcinoma	
HIT000036226	NPD002	acyl-CoenzymeAdehydrogenasefamily,member9;acyl-CoA	Lung, small cell carcinoma	1- and 2-Methylnaphthalene degradation,Bile acid biosynthesis
HIT000031933	PSMA7	Proteasomesubunitalphatype7(EC3.4.25.1)(Proteasomesubunit	Lung, small cell carcinoma	Proteasome
HIT000005469	FAM26B	Conservedhypotheticalprotein,completecds.	Mammary gland	
HIT000001825		Ubiquitouslyexpressedtranscriptfamilyprotein,completecds.	normal pituitary	
HIT000004129	PARVA	Alpha-parvin(Calponin-likeintegrin-linkedkinasebinding	ovarian cancer	Focal adhesion
HIT000032059	C1orf91, RP4-622L5,RP4-622L5.3	novel protein, similar to AASL548	Ovary, adenocarcinoma	
HIT000035989	TREB5	Xboxbindingprotein-1(XBP-1)(TREB5protein),completecds.	Ovary, adenocarcinoma	
HIT000006270	ACTR8	Actin-relatedprotein8,completecds.	ovary, tumor tissue	
HIT000005768	RAB3GAP1	Conservedhypotheticalprotein,partialcds.	ovary, tumor tissue	
HIT000006553	VPS11	Vacuolarproteinsorting11(hVPS11)(PP3476),completecds.	ovary, tumor tissue	
HIT000010702	ZDHHC12	Zn-finger,DHHCtypedomaincontainingprotein,completecds.	ovary, tumor tissue	
HIT000006231	ZNF447	Zn-finger,C2H2typedomaincontainingprotein,completecds.	ovary, tumor tissue	
HIT000001936	ARBP	SimilartoBrainprotein44-likeprotein(Apoptosis-regulating	Pituitary	
HIT000001783	RPS27L	ribosomalproteinS27-likeprotein;40SribosomalproteinS27	Pituitary	Ribosome
HIT000017381	AOC1	Amiloride-sensitiveamineoxidase(copper-containing)precursor	Placenta	Tryptophan metabolism beta-Alanine metabolism Phenylalanine metabolism

				Histidine metabolism Glycine, serine and threonine metabolism Alkaloid biosynthesis II Arginine and proline metabolism Tyrosine metabolism
HIT000013380	VIM	SimilartoVimentin,partialcds.	Placenta	
HIT000017327		Non-protein-codingtranscript,completesequence.	Placenta	
HIT000030493	ELE1	Nuclearreceptorcoactivator4(NCoA-4)(70kDaandrogenreceptor	Placenta, choriocarcinoma	
HIT000032951	RBM3	RNA-bindingregionRNP-1(RNArecognitionmotif)domain	Placenta, choriocarcinoma	
HIT000029861	SERS	Seryl-tRNAsynthetase(EC6.1.1.11)(Serine--tRNAligase)(SerRS),	Placenta, choriocarcinoma	
HIT000032280	ALEX3	ALEX3protein;armproteinlostinepithelialcancers,X	Skin, melanotic melanoma, high MDR.	
HIT000036559	ADPRHL2	ADP-ribosylglycohydrolasefamilyprotein,completecds.	Skin, melanotic melanoma.	
HIT000032646	OBFC2B,MGC2731	Conservedhypotheticalprotein,completecds.	Skin, melanotic melanoma.	
HIT000036906	PCCX1	CpGbindingprotein(ProteincontainingPHDfingerandCXXCdomain	Skin, melanotic melanoma.	
HIT000035319	TFIP11	tuftelininteractingprotein11(Homosapiens),completecds.	Skin, melanotic melanoma.	
HIT000013648	MXD3	SimilartoMAXdimerizationprotein3(Homosapiens),completecds.	small intestine	
HIT000014896	FLJ00187,MSZF13	SimilartoMszf13(Fragment),completecds.	Spleen	
HIT000014696		G-proteinbetaWD-40repeatcontainingprotein,partialcds.	Testis	
HIT000022539		Hypotheticalprotein,completecds.	Testis	
HIT000007189	DAK	Dakkinasedomaincontainingprotein,completecds.	thyroid gland	
HIT000012628	NDEL1	nudEnucleardistributiongeneEhomologlike1(A.nidulans);	Tongue	
HIT000012948	USP39	SimilartoU4/U6.U5tri-snRNP-associated65kDaprotein,partial	Tongue	

HIT000012839		Ankyrinrepeatcontainingprotein,completecds.	Tongue	
HIT000017137		Conservedhypotheticalprotein,partialcds.	Tongue	
HIT000023156	KCNMA1	Hypotheticalprotein,completecds.	Uterus	
HIT000040041	DYNC1LI2,DNCLI2	Dynein,cytoplasmic,lightintermediatepolypeptide2,complete	Uterus, leiomyosarcoma	
HIT000039050	YWHAG	14-3-3proteingamma(ProteinkinaseCinhibitorprotein-1)	Uterus, leiomyosarcoma	Cell cycle
HIT000004823		Hypotheticalprotein,completecds.	whole embryo, mainly head	
HIT000007314		Hypotheticalprotein,completecds.	whole embryo, mainly head	

Supplementary Table 7.4a DEFA1 predicted genes that matched co-expression data

H-inv ID	Gene symbol	Gene Description	Pathway	Tissue origin
HIT000035146	CCND2	G1/S-specific cyclin D2, partial cds.	hsa04110 Cell cycle, hsa04310 Wnt signaling pathway, hsa04510 Focal adhesion,hsa04630 Jak	Bone marrow, chronic myelogenous leukemia
HIT000042211	KIAA0065, MSZF68, ZNF33A	Similar to Mszf68 (Fragment), partial cds.		Brain
HIT000037044	MIZ1, PIAS2	Protein inhibitor of activated STAT X isoform alpha		Brain, glioblastoma
HIT000035490	MARS	methionine-tRNA synthetase; methionine tRNA ligase; methionyl-tRNA	hsa00271 Methionine metabolism, hsa00450 Selenoamino acid metabolism, hsa00970 Aminoacyl-tRNA synthetases	Brain, neuroblastoma
HIT000033138	VSIG2	Immunoglobulin subtype domain containing protein, complete cds.		Colon ,adenocarcinoma
HIT000030627	MEA1, MEA	Male-enhanced antigen-1 (Mea-1), complete cds.		Eye, retinoblastoma
HIT000003137	CNOT2	CCR4-NOT transcription complex, subunit 2; NOT2 (negative)		Ileal mucosa

HIT000040168	UGT2B11	UDP-glucuronosyltransferase 2B4 precursor, microsomal	hsa00040 Pentose and glucuronate interconversions hsa00150 Androgen and estrogen metabolism, hsa00500 Starch and sucrose metabolism, hsa00860 Porphyrin and chlorophyll metabolism	Liver
HIT000030833	DDX23	DEAD/DEAH box helicase domain containing protein, partial	hsa00500 Starch and sucrose metabolism, hsa00790 Folate biosynthesis	Lung,smallcellcarcinoma
HIT000033214	ARB2, ARRB2	Beta-arrestin 2 (Arrestin, beta 2), partial cds.	MAPK signalling	Muscle,rhabdomyosarcoma
HIT000032247	FKBP12, FKBP1A	FK506-binding protein 1A (EC 5.2.1.8)	mTOR signalling pathway	Placenta,choriocarcinoma
HIT000030153	TMED9	Emp24/gp25L/p24 family protein, complete cds.		Skin,melanoticmelanoma.
HIT000030145	Y2, PSMB8	Similar to Proteasome subunit beta type 8 precursor (EC 3.4.25.1)		Skin,melanoticmelanoma.
HIT000030885		Caspase-1 precursor, p45 family protein, complete cds.		Skin,melanoticmelanoma.
HIT000038574	LABH2, ABHD2	Abhydrolase domain containing protein 2 (Protein PHPS1-2),		Testis
HIT000037134	MARCH5	Zn-finger, RING domain containing protein, complete cds.		Uterus,leiomyosarcoma
HIT000032576	RSU	Ras protein 1 (Rsu-1) (RSP-1), complete cds.	n/a	Brain,primitiveneuroectodermal

Supplementary Table 7.4b: Gene hits from DEFA1 promoter model scan that did not match co-expressed gene data for DEFA1, DEFA3

H-inv ID	Gene name	Gene Description	Pathway	1..1.1.1.2 Tissue
HIT000014970	SSB3	SPRY domain-containing SOCS box protein SSB-3 (Homo sapiens)		adiposetissue
HIT000041181	NA	Pleckstrin putative G-protein interacting domain containing		Bonemarrow, chronicmyelogenousleukemia
HIT000000028	KIAA0322, NEDL1, HECW1	NEDD4-like ubiquitin ligase 1, complete cds.		brain
HIT000000210	KIAA0494,	Calcium-binding EF-hand domain containing protein, complete cds.		brain
HIT000000384	KIAA0659, C11orf11	Lipase, class 3 family protein, partial cds.		brain
HIT000025067	DKFZp564C047, STAM2	Signal transducing adaptor molecule 2; STAM-like protein	hsa04630 Jak-STAT signaling pathway	brain
HIT000040216	NA	Chaperonin Cpn60		Brain, hippocampus
HIT000029388	NICE-3, C1orf43	NICE-3 protein (Homo sapiens), complete cds.		Cervix, carcinoma
HIT000008039	PSARL	Rhomboid-like protein family protein, complete cds.		colon
HIT000030636	CASM,LSM1	U6 snRNA-associated Sm-like protein LSm1 (Small nuclear		Eye,retinoblastoma
HIT000028516	DKFZp686K0367, ZNF-kaiso,ZBTB33	Kaiso (Homo sapiens), complete cds.		humanendometriumcarcinomacelline
HIT000009149	NA	Prolyl 4-hydroxylase, alpha subunit family protein, complete cds.		humansmallintestine
HIT000033091	LZIC	Conserved hypothetical protein, complete cds.		Kidney,renalcelladenocarcinoma

HIT000029686	SUCLG1,SUCLG1	Succinyl-CoA ligase (GDP-forming) alpha-chain, mitochondrial	hsa00020 Citrate cycle (TCA cycle),hsa00640 Propanoate metabolism	Lung,smallcellcarcinoma
HIT000031367	ACY1,ACY1	Aminoacylase-1 (EC 3.5.1.14) (N-acyl-L-amino-acid amidohydrolase)	hsa00220 Urea cycle and metabolism of amino groups	Lung,smallcellcarcinoma
HIT000029034	DKFZp762G014,KIAA1172, SFRS15	CTD-binding SR-like protein RA4 (Fragment). Splice isoform 2,		melanoma(MeWocellline)
HIT000003713	RBM28	RNA-binding region RNP-1 (RNA recognition motif) domain		ovariancancer
HIT000004151	NA	Amino acid/polyamine transporter, family II protein, complete cds.		ovariancancer
HIT000004194	FLJ10858,NEIL3	DNA glycosylase hFPG2 (Homo sapiens), complete cds.		ovariancancer
HIT000037144	EDF1,NA	endothelial differentiation-related factor 1 isoform alpha;		Pancreas,epithelioidcarcinoma
HIT000004532	LARP6	RNA-binding protein Lupus Lal domain containing protein, complete		placenta
HIT000030445	MDHA,MDH1	Malate dehydrogenase, cytoplasmic (EC 1.1.1.37), partial cds.		Placenta,choriocarcinoma
HIT000034068	ECHS1,ECHS1	Enoyl-CoA hydratase, mitochondrial precursor (EC 4.2.1.17) (Short		Placenta,choriocarcinoma
HIT000038819	ZNF183	Zn-finger, C-x8-C-x5-C-x3-H type domain containing protein,		Placenta,choriocarcinoma
HIT000035139	GPIP4,PIP	Prolactin-inducible protein precursor (Secretory actin-binding		Prostate
HIT000039309	TCF5,CEBPB	CCAAT/enhancer binding protein beta (C/EBP beta)		Skin,melanoticmelanoma,highMDR.

HIT000029801	RING5,SLC39A7	Histidine-rich membrane protein Ke4, partial cds.		Skin,melanoticmelanoma.
HIT000014971	ZNF414	Zn-finger, C2H2 type domain containing protein, complete cds.		
HIT000000974	KIAA1258,GDA	Guanine deaminase (EC 3.5.4.3) (Guanase) (Guanine aminase)		adiposetissue
HIT000000279	KIAA0562,	Similar to Glycine-, glutamate-,		brain
HIT000042167	KIAA1293,KIAA1293, FDPS	Farnesyl pyrophosphate synthetase (FPP synthetase) (FPS) (Farnesyl		brain
HIT000042279	KIAA0142,KIAA0142, ARHGEF7	DEFINITION: Rho guanine nucleotide exchange factor 7 (PAK-interacting exchange		brain
HIT000041526	ATP6V0A1	V-type ATPase, 116 kDa subunit family protein, complete cds.		brain
HIT000035869	MRPS10,MRPS10	Mitochondrial 28S ribosomal protein S10 (MRP-S10) (MSTP040),		Brain,adult,6pooledwholebrains
HIT000041711	OPTN,OPTN,NRP	optineurin; tumor necrosis factor alpha-inducible cellular protein		Brain,primitiveneuroectodermal
HIT000015206	UBAP1,NA	ubiquitin associated protein (Homo sapiens), complete cds.		Cervix,carcinoma
HIT000038183	VAMP5,NA	vesicle-associated membrane protein 5 (myobrevin) (Homo sapiens),		humanlung
HIT000035479	NA	PPR repeat containing protein, complete cds.		Lung
HIT000030704	HIG2, NA	Hypoxia-inducible protein 2, complete cds.		Lung,smallcellcarcinoma
HIT000030303	YIF1	Hrf1 family protein, complete cds.		Lung,smallcellcarcinoma
HIT000032951	RBM3	RNA-binding region RNP-1 (RNA recognition motif) domain		Placenta,choriocarcinoma
HIT000017533	STEAP2	NADP oxidoreductase, coenzyme F420-dependent family protein,		Placenta,choriocarcinoma
HIT000039707	ORF1-FL49	Molluscan rhodopsin C-terminal tail family protein, complete cds.		prostate
HIT000021780	NA	Questionable transcript.		Skin,melanoticmelanoma,highMDR.
HIT000007696	FLJ00011,PDZK7	DEFINITION: PDZ/DHR/GLGF domain containing protein, complete cds.		smallintestine

HIT000038678	APRIL,ANP32B	Acidic leucine-rich nuclear phosphoprotein 32 family member B		spleen
Model hits which are hypothetical protein				Testis,embryonalcarcinoma
HIT000029004	DKFZp547B1713,	Conserved hypothetical protein, complete cds.		
HIT000039620	NA	Conserved hypothetical protein, complete cds.		brain
HIT000028000	DKFZp451M2119,	Hypothetical protein, complete cds.		Brain,hippocampus
HIT000033703	NA	Conserved hypothetical protein, complete cds.		humanspinalcord
HIT000032617	C14orf160	Conserved hypothetical protein, complete cds.		Lung,smallcellcarcinoma
HIT000020033	NA	Hypothetical protein, complete cds.		Placenta,choriocarcinoma
HIT000013900	PACRG	Conserved hypothetical protein, complete cds.		substantianigra
HIT000021200	NA	Conserved hypothetical protein, partial cds.		testis

Supplementary Table 7.5a: Alpha defensin1 predicted genes clustered based on GO biological function

Percentage: The number of genes that have a particular GO term / total number of genes that have a GO term

Clustering based on GO level 4			
GO ID	**GO Biological function**	**Gene name**	**Percentage**
GO:0044237	cellular metabolism	NDUFS5 MARCH5 TAF11 DPM1 UGT2B11 MYST2 CHD2 H1F0 PECI CNOT2 DDX23 MARS INS	48.15
GO:0044238	primary metabolism	MARCH5 TAF11 DPM1 UGT2B11 MYST2 CHD2 H1F0 CPNE6 PECI CNOT2 DDX23 MARS INS	48.15
GO:0006810	Transport	MIP FTL ATP5S CPNE6 SRPR SEC5L1	22.22
GO:0051234	establishment of localization	MIP FTL ATP5S CPNE6 SRPR SEC5L1	22.22
GO:0051244	regulation of cellular physiological process	TAF11 MYST2 CCND2 CHD2 CNOT2 CCNI	22.22
GO:0043170	macromolecule metabolism	MARCH5 DPM1 H1F0 MARS INS	18.52
GO:0043283	biopolymer metabolism	MYST2 CHD2 H1F0 DDX23 MARS	18.52
GO:0007165	signal transduction	APBB1IP CX3CL1 FMOD INS	14.81
GO:0019222	regulation of metabolism	TAF11 MYST2 CHD2 CNOT2	14.81
GO:0007049	cell cycle	DCTN3 CCND2 CCNI	11.11
GO:0007267	cell-cell signaling	MIP CPNE6 INS	11.11
GO:0009605	response to external stimulus	UGT2B11 MIP CX3CL1	11.11
GO:0007155	cell adhesion	TMEM8 CX3CL1	7.41
GO:0008104	protein localization	SRPR SEC5L1	7.41
GO:0009058	Biosynthesis	DPM1 MARS	7.41
GO:0009887	Organogenesis	CPNE6 KRT5	7.41
GO:0016043	cell organization and biogenesis	CHD2 H1F0	7.41
GO:0045045	secretory pathway	SRPR SEC5L1	7.41
GO:0050877	neurophysiological process	MIP CPNE6	7.41
GO:0051301	cell division	DCTN3 CCND2	7.41
GO:0001775	cell activation	CX3CL1	3.7
GO:0006928	cell motility	CX3CL1	3.7
GO:0006950	response to stress	CX3CL1	3.7
GO:0006955	immune response	CX3CL1	3.7

Supplementary Table 7.5b: Alpha defensin1 predicted genes clustered based on molecular function.

Percentage: (The number of genes that have a particular GO term / total number of genes in the dataset that have GO annotation)

GO ID	GO molecular function	Gene Names	Percentage
GO:0003677	DNA binding	CHD1L TAF11 MYST2 CHD2 NCL H1F0	23.08
GO:0017076	purine nucleotide binding	CHD1L CHD2 DDX23 SRPR MARS NUBP2	23.08
GO:0003723	RNA binding	NCL DDX23 AKAP1 SRPR MARS	19.23
GO:0016817	hydrolase activity, acting on acid anhydrides	CHD1L CHD2 DDX23 SRPR	15.38
GO:0043169	cation binding	MARCH5 MYST2 FTL CPNE6	15.38
GO:0046872	metal ion binding	MARCH5 MYST2 FTL CPNE6	15.38
GO:0008026	ATP-dependent helicase activity	CHD1L CHD2 DDX23	11.54

Supplementary Table 7.6a: DEFA5 predicted genes that matched co-expressed genes classified based on GO biological function

Percentage: The number of genes that have a particular GO term / total number of genes in the dataset that have a GO annotation

Biological Function	Genes	No.of genes	Percentage
cellular metabolism	CLG4A KIAA0175 MRPL4 PRIM1 AP2B1 SLUG ATPIF1 ILF2 PRDX2 TXNRD1 MBD1 TTF DENR TAF11 SCAND2 ISOT HER3 LTB4DH TOMM34 TSSC3 HSPA14 CEV14 ATPAF1 KIAA0106 SPC18 APOL1 KIAA0517 TYRO10 SAP61 MRPS36 PFKL NMP238 MRPS2 NUDT5 KIAA1667 MKI67IP SAP114 AUH MID1 UGP2 TRIM38 NEC2 MNK1 PLAGL2 UBQLN1 RPP40 DHLAG GM2A RNMTL1 KIAA0091 TRX1 CBX6 HOX1G NAALAD1 PIM2	55	67.9
primary metabolism	CLG4A KIAA0175 MRPL4 PRIM1 AP2B1 SLUG ATPIF1 ILF2 MBD1 TTF DENR TAF11 SCAND2 ISOT HER3 LTB4DH TOMM34 TSSC3 HSPA14 CEV14 ATPAF1 KIAA0106 SPC18 APOL1 KIAA0517 TYRO10 SAP61 MRPS36 PFKL NMP238 MRPS2 NUDT5 KIAA1667 MKI67IP SAP114 AUH MID1 UGP2 TRIM38 NEC2 MNK1 PLAGL2 UBQLN1 RPP40 DHLAG GM2A RNMTL1 KIAA0091 TRX1 CBX6 HOX1G NAALAD1 PIM2	53	65.43
macromolecule metabolism	CLG4A KIAA0175 MRPL4 PRIM1 AP2B1 DENR ISOT HER3 TOMM34 HSPA14 ATPAF1 SPC18 APOL1 KIAA0517 TYRO10 SAP61 MRPS36 PFKL NMP238 MRPS2 NUDT5 KIAA1667 MKI67IP SAP114 AUH MID1 UGP2 TRIM38 NEC2 MNK1 UBQLN1 RPP40 DHLAG RNMTL1 KIAA0091 TRX1 CBX6 NAALAD1 PIM2	39	48.15
regulation of cellular physiological process	SLUG ATPIF1 ILF2 PRDX2 CCNB MBD1 TTF TAF11 SCAND2 NMP238 MNK1 PLAGL2 AIF1 DHLAG TRX1 CBX6 HOX1G	17	20.99
establishment of localization	STX6 AP2B1 TXNRD1 STAU2 TTF TOMM34 APOL1 SATT MSTP028 BAP29 SLC17A5 KIAA1667 CTRP6 ARFL3 NESH DHLAG COL9A3	17	20.99
Transport	STX6 AP2B1 TXNRD1 STAU2 TTF TOMM34 APOL1 SATT MSTP028 BAP29 SLC17A5 KIAA1667 CTRP6 ARFL3 DHLAG COL9A3	16	19.75
regulation of metabolism	SLUG ATPIF1 ILF2 MBD1 TTF TAF11 SCAND2 NMP238 MNK1 PLAGL2 TRX1 CBX6 HOX1G	13	16.05
signal transduction	TXNRD1 TTF HER3 TYRO10 DRG2 ARFL3 TRIM38 MNK1 DHLAG PAK1IP1	10	12.35
immune response	HLA-DMA ILF2 G10P1 TTF IFI30 AIF1 DHLAG	7	8.64
response to pest, pathogen or parasite	HLA-DMA AIF1 DHLAG	3	3.7
hemopoietic or lymphoid organ development	TTF DHLAG TRX1	3	3.7
lymphocyte differentiation	TTF DHLAG	2	2.47

Supplementary Table 7.6b: DEFA5 novel predicted genes classified based on GO biological function

Percentage: The number of genes that have a particular GO term / total number of genes in the dataset that have a GO annotation

Biological Function	Genes	No. of genes	Percentage
cellular metabolism	PCCX1 NPD002 VPS11 TRIM9 EFCBP1 ELE1 RPS27L RBM3 PSMA7 MXD3 H2AFX EPM2A USP39	13	65
primary metabolism	PCCX1 VPS11 TRIM9 ELE1 RPS27L RBM3 PSMA7 MXD3 H2AFX EPM2A USP39	11	55
macromolecule metabolism	VPS11 TRIM9 RPS27L RBM3 PSMA7 H2AFX EPM2A USP39	8	40
establishment of localization	NPD002 VPS11 CX3CL1 VIM TFIP11 KCNMA1	6	30
Transport	NPD002 VPS11 VIM TFIP11 KCNMA1	5	25
signal transduction	YWHAG TENC1 CX3CL1 ELE1 PDZK2	5	25
immune response	CX3CL1	1	5
response to pest, pathogen or parasite	CX3CL1	1	5

Supplementary Table 7.7: Common regulatory elements found across the predicted set of genes from DEAF1 and DEFA5 models.

The transcription factors here are predicted using FATIGO+ (*http://babelomics.bioinfo.cipf.es/fatigoplus/cgi-bin/fatigoplus.cgi*)

Transcription factors DEFA5 Matched gene hits	Genes	No. of genes
HNF-1	STX6 CLG4A HLA-DMA AK000471 KIAA0175 MRPL4 AP2B1 WBSCR17 SLUG ATPIF1 RDH1 UCC1 ILF2 RPMS13 TMEM30A RCBTB1 PRDX2 HDHD1A G10P1 BC093018 CCNB STAU2 MBD1 TTF DENR TAF11 SCAND2 SPATS2 ISOT LGALS4 TIN2 HER3 PIGT LTB4DH TOMM34 TSSC3 HSPA14 HRBL CEV14 PODXL ATPAF1 KIAA0106 SPC18 IFI30 APOL1 SATT KIAA0517 TYRO10 MSTP028 KIAA1745 BAP29 DRG2 SAP61 DDX49 TMSB10 BGN MRPS36 COPS7A DERP7 AK096925 PFKL NMP238 KBTBD6 MRPS2 SLC17A5 AF454939 NUDT5 BRMS1L KIAA1667 MKI67IP MRS1 SAP114 BC051849 CTRP6 DDX31 GFAP AUH AK026266 ARFL3 MID1 UGP2 SNAI1 RNUT1 TRIM38 NEC2 MNK1 PLAGL2 UBQLN1 RPP40 HN1 AIF1 KIAA1624 NESH LABH2 DHLAG TMEM35 MRPL46 CCDC12 TMEM33 RNMTL1 CKLFSF6 KIAA0091 TRX1 CBX6 MAP2K1IP1 HOX1G CLDN2 NAALAD1 COL9A3 PIM2 PAK1IP1 KIAA0638	112
CDX	STX6 HLA-DMA KIAA0175 MRPL4 AP2B1 WBSCR17 SLUG TBC1D17 ATPIF1 RDH1 UCC1 ILF2 RPMS13 TMEM30A RCBTB1 PRDX2 HDHD1A G10P1 BC093018 CCNB STAU2 MBD1 TTF DENR TAF11 SCAND2 SPATS2 ISOT LGALS4 TIN2 HER3 PIGT LTB4DH TOMM34 TSSC3 HSPA14 HRBL CEV14 PODXL ATPAF1 KIAA0106 SPC18 IFI30 APOL1 SATT KIAA0517 TYRO10 MSTP028 KIAA1745 BAP29 DRG2 SAP61 DDX49 TMSB10 BGN MRPS36 COPS7A DERP7 AK096925 NMP238 KBTBD6 MRPS2 SLC17A5 AF454939 NUDT5 BRMS1L MKI67IP MRS1 SAP114 BC051849 CTRP6 DDX31 GFAP AUH AK026266 ARFL3 MID1 UGP2 SNAI1 RNUT1 TRIM38 NEC2 MNK1 PLAGL2 UBQLN1 RPP40 HN1 AIF1 KIAA1624 NESH LABH2 DHLAG TMEM35 MRPL46 CCDC12 TMEM33 RNMTL1 CKLFSF6 KIAA0091 TRX1 CBX6 MAP2K1IP1 HOX1G CLDN2 NAALAD1 COL9A3 PIM2 PAK1IP1 KIAA0638	109
Nkx2-5	STX6 CLG4A HLA-DMA AK000471 KIAA0175 MRPL4 AP2B1 WBSCR17 SLUG ATPIF1	108

	RDH1 UCC1 ILF2 RPMS13 TMEM30A RCBTB1 PRDX2 HDHD1A G10P1 BC093018 CCNB STAU2 TTF DENR TAF11 SCAND2 SPATS2 ISOT LGALS4 TIN2 HER3 PIGT LTB4DH TOMM34 TSSC3 HSPA14 HRBL CEV14 PODXL ATPAF1 KIAA0106 SPC18 IFI30 APOL1 SATT KIAA0517 TYRO10 MSTP028 KIAA1745 BAP29 DRG2 SAP61 DDX49 TMSB10 MRPS36 COPS7A DERP7 AK096925 PFKL NMP238 KBTBD6 MRPS2 SLC17A5 AF454939 NUDT5 BRMS1L KIAA1667 MKI67IP SAP114 BC051849 CTRP6 DDX31 GFAP AUH AK026266 ARFL3 MID1 UGP2 SNAI1 RNUT1 TRIM38 NEC2 MNK1 PLAGL2 UBQLN1 RPP40 HN1 AIF1 KIAA1624 NESH LABH2 DHLAG TMEM35 MRPL46 CCDC12 TMEM33 RNMTL1 CKLFSF6 KIAA0091 TRX1 CBX6 MAP2K1IP1 CLDN2 NAALAD1 COL9A3 PIM2 PAK1IP1 KIAA0638	
GATA-4	STX6 CLG4A HLA-DMA AK000471 KIAA0175 MRPL4 AP2B1 WBSCR17 SLUG TBC1D17 ATPIF1 RDH1 UCC1 ILF2 RPMS13 TMEM30A RCBTB1 PRDX2 HDHD1A G10P1 BC093018 CCNB STAU2 MBD1 TTF DENR TAF11 SCAND2 SPATS2 ISOT LGALS4 HER3 PIGT LTB4DH TOMM34 TSSC3 HSPA14 HRBL CEV14 PODXL ATPAF1 KIAA0106 SPC18 APOL1 SATT KIAA0517 TYRO10 MSTP028 KIAA1745 BAP29 DRG2 SAP61 DDX49 TMSB10 MRPS36 COPS7A DERP7 AK096925 NMP238 MRPS2 SLC17A5 AF454939 NUDT5 BRMS1L KIAA1667 MKI67IP MRS1 SAP114 BC051849 CTRP6 DDX31 GFAP AUH AK026266 ARFL3 MID1 UGP2 SNAI1 RNUT1 TRIM38 NEC2 MNK1 PLAGL2 UBQLN1 RPP40 HN1 AIF1 KIAA1624 NESH LABH2 DHLAG TMEM35 MRPL46 CCDC12 TMEM33 RNMTL1 CKLFSF6 KIAA0091 TRX1 CBX6 MAP2K1IP1 HOX1G CLDN2 NAALAD1 COL9A3 PIM2 PAK1IP1 KIAA0638	108
LXR, PXR, CAR, COUP, RAR	STX6 CLG4A HLA-DMA AK000471 KIAA0175 MRPL4 AP2B1 WBSCR17 TBC1D17 RDH1 UCC1 ILF2 RPMS13 TMEM30A RCBTB1 PRDX2 HDHD1A G10P1 CCNB STAU2 MBD1 TTF DENR TAF11 SCAND2 SPATS2 ISOT LGALS4 TIN2 HER3 PIGT LTB4DH TOMM34 TSSC3 HSPA14 HRBL CEV14 PODXL ATPAF1 KIAA0106 SPC18 IFI30 APOL1 SATT KIAA0517 TYRO10 MSTP028 KIAA1745 BAP29 DRG2 SAP61 DDX49 TMSB10 BGN MRPS36 COPS7A DERP7 PFKL NMP238 KBTBD6 MRPS2 SLC17A5 AF454939 NUDT5 BRMS1L KIAA1667 MKI67IP MRS1 SAP114 BC051849 CTRP6 DDX31 GFAP AUH AK026266 ARFL3 UGP2 SNAI1 RNUT1 TRIM38 MNK1 PLAGL2 UBQLN1 RPP40 HN1 AIF1 KIAA1624 NESH LABH2 DHLAG TMEM35 MRPL46 CCDC12 TMEM33 RNMTL1 CKLFSF6 KIAA0091 TRX1 CBX6 HOX1G CLDN2 NAALAD1 COL9A3 PIM2 KIAA0638	105
Oct-1	STX6 CLG4A HLA-DMA AK000471 KIAA0175 MRPL4 AP2B1 WBSCR17 SLUG TBC1D17 ATPIF1 RDH1 UCC1 ILF2 RPMS13 TMEM30A HDHD1A G10P1 BC093018 CCNB STAU2 TTF	104

	DENR TAF11 SCAND2 SPATS2 ISOT TIN2 HER3 PIGT LTB4DH TOMM34 TSSC3 HSPA14 HRBL CEV14 PODXL ATPAF1 KIAA0106 SPC18 IFI30 APOL1 SATT KIAA0517 TYRO10 MSTP028 KIAA1745 BAP29 DRG2 SAP61 DDX49 TMSB10 MRPS36 COPS7A DERP7 AK096925 NMP238 KBTBD6 MRPS2 SLC17A5 AF454939 NUDT5 BRMS1L KIAA1667 MKI67IP MRS1 SAP114 BC051849 CTRP6 DDX31 GFAP AUH ARFL3 MID1 UGP2 SNAI1 RNUT1 TRIM38 NEC2 MNK1 PLAGL2 UBQLN1 RPP40 HN1 AIF1 KIAA1624 NESH LABH2 DHLAG TMEM35 MRPL46 TMEM33 RNMTL1 CKLFSF6 KIAA0091 TRX1 CBX6 MAP2K1IP1 HOX1G CLDN2 NAALAD1 PIM2 PAK1IP1 KIAA0638	
Cdc5	STX6 CLG4A HLA-DMA AK000471 KIAA0175 AP2B1 WBSCR17 SLUG TBC1D17 ATPIF1 RDH1 UCC1 RPMS13 TMEM30A RCBTB1 HDHD1A G10P1 BC093018 CCNB STAU2 MBD1 TTF DENR TAF11 SCAND2 SPATS2 ISOT LGALS4 TIN2 HER3 PIGT LTB4DH TOMM34 HSPA14 HRBL CEV14 PODXL ATPAF1 KIAA0106 SPC18 IFI30 APOL1 SATT KIAA0517 TYRO10 MSTP028 KIAA1745 BAP29 DRG2 SAP61 TMSB10 MRPS36 COPS7A DERP7 AK096925 NMP238 KBTBD6 MRPS2 SLC17A5 AF454939 NUDT5 BRMS1L KIAA1667 MKI67IP MRS1 BC051849 CTRP6 DDX31 AUH AK026266 ARFL3 MID1 UGP2 SNAI1 RNUT1 TRIM38 NEC2 MNK1 PLAGL2 UBQLN1 RPP40 HN1 KIAA1624 NESH LABH2 DHLAG TMEM35 MRPL46 TMEM33 RNMTL1 CKLFSF6 KIAA0091 TRX1 CBX6 MAP2K1IP1 HOX1G CLDN2 NAALAD1 COL9A3 PIM2 PAK1IP1 KIAA0638	102
NF-kappaB	STX6 CLG4A HLA-DMA AK000471 KIAA0175 MRPL4 AP2B1 WBSCR17 SLUG TBC1D17 ATPIF1 RDH1 UCC1 ILF2 RPMS13 TMEM30A RCBTB1 PRDX2 HDHD1A G10P1 BC093018 CCNB STAU2 MBD1 TTF DENR TAF11 SCAND2 SPATS2 ISOT LGALS4 TIN2 PIGT LTB4DH TOMM34 TSSC3 HSPA14 CEV14 PODXL ATPAF1 KIAA0106 IFI30 APOL1 SATT KIAA0517 MSTP028 KIAA1745 BAP29 DRG2 SAP61 TMSB10 BGN MRPS36 COPS7A DERP7 AK096925 PFKL NMP238 KBTBD6 MRPS2 SLC17A5 AF454939 BRMS1L KIAA1667 MKI67IP MRS1 SAP114 BC051849 CTRP6 DDX31 GFAP AUH AK026266 MID1 UGP2 SNAI1 RNUT1 TRIM38 NEC2 MNK1 UBQLN1 RPP40 HN1 AIF1 KIAA1624 NESH LABH2 DHLAG MRPL46 CCDC12 TMEM33 RNMTL1 CKLFSF6 CBX6 MAP2K1IP1 HOX1G CLDN2 NAALAD1 COL9A3 PIM2 PAK1IP1 KIAA0638	102

Unmatched gene hits		
LXR, PXR, CAR, COUP, RAR	PCCX1 YWHAG NPD002 VPS11 TENC1 TRIM9 FAM45A CX3CL1 AOC1 EFCBP1 ADPRHL2 ALEX3 VIM ELE1 PDZK2 ZMYND13 TFIP11 RPS27L RBM3 PSMA7 MXD3 FAM26B ACTR8 H2AFX DNCLI2 EPM2A USP39 NDEL1 LRRN6A KCNMA1 ZDHHC12 TM4SF1	32
CDX	PCCX1 YWHAG NPD002 VPS11 TENC1 TRIM9 FAM45A CX3CL1 AOC1 EFCBP1 ADPRHL2 ALEX3 VIM ELE1 PDZK2 ZMYND13 TFIP11 RPS27L RBM3 PSMA7 MXD3 FAM26B ACTR8 H2AFX DNCLI2 EPM2A USP39 NDEL1 LRRN6A KCNMA1 ZDHHC12 TM4SF1	32
HNF-1	PCCX1 YWHAG NPD002 VPS11 TENC1 TRIM9 FAM45A CX3CL1 AOC1 EFCBP1 ADPRHL2 ALEX3 VIM ELE1 PDZK2 ZMYND13 TFIP11 RPS27L RBM3 PSMA7 MXD3 FAM26B ACTR8 H2AFX DNCLI2 EPM2A USP39 NDEL1 LRRN6A KCNMA1 ZDHHC12 TM4SF1	32
Nkx2-5	PCCX1 YWHAG NPD002 VPS11 TENC1 TRIM9 FAM45A CX3CL1 AOC1 EFCBP1 ADPRHL2 ALEX3 VIM ELE1 PDZK2 ZMYND13 TFIP11 RPS27L RBM3 MXD3 FAM26B ACTR8 H2AFX DNCLI2 EPM2A USP39 NDEL1 LRRN6A KCNMA1 ZDHHC12 TM4SF1	31
GATA-4	PCCX1 YWHAG NPD002 VPS11 TENC1 TRIM9 FAM45A CX3CL1 AOC1 EFCBP1 ADPRHL2 ALEX3 VIM ELE1 PDZK2 ZMYND13 TFIP11 RPS27L RBM3 MXD3 FAM26B ACTR8 H2AFX DNCLI2 EPM2A USP39 NDEL1 LRRN6A KCNMA1 ZDHHC12 TM4SF1	31
Oct-1	PCCX1 YWHAG NPD002 VPS11 TENC1 TRIM9 CX3CL1 AOC1 EFCBP1 ADPRHL2 ALEX3 VIM ELE1 PDZK2 ZMYND13 TFIP11 RPS27L RBM3 PSMA7 MXD3 FAM26B ACTR8 H2AFX DNCLI2 EPM2A USP39 NDEL1 LRRN6A KCNMA1 ZDHHC12 TM4SF1	31
NF-kappaB	PCCX1 NPD002 VPS11 TENC1 TRIM9 FAM45A CX3CL1 AOC1 EFCBP1 ADPRHL2 ALEX3 VIM ELE1 PDZK2 ZMYND13 TFIP11 RPS27L RBM3 MXD3 FAM26B ACTR8 H2AFX DNCLI2 EPM2A USP39 NDEL1 LRRN6A KCNMA1 ZDHHC12 TM4SF1	30
DEFA1		
Matched gene hits		
PPAR direct repeat 1	RBM3 ATP6V0A1 STEAP2 LZIC ARHGEF7 SFRS15 PACRG PSARL RBM28 FDPS	10
Nkx2-5	RBM3 ATP6V0A1 STEAP2 LZIC ARHGEF7 SFRS15 PACRG PSARL RBM28 FDPS	10
Cdc5	RBM3 ATP6V0A1 STEAP2 LZIC ARHGEF7 SFRS15 PACRG PSARL RBM28 FDPS	10
GATA-4		10

	RBM3 ATP6V0A1 STEAP2 LZIC ARHGEF7 SFRS15 PACRG PSARL RBM28 FDPS	
CDX	RBM3 ATP6V0A1 STEAP2 LZIC ARHGEF7 SFRS15 PACRG PSARL RBM28 FDPS	10
HNF-1	RBM3 ATP6V0A1 STEAP2 LZIC ARHGEF7 SFRS15 PACRG PSARL RBM28 FDPS	10
NF-kappaB	RBM3 ATP6V0A1 STEAP2 LZIC ARHGEF7 SFRS15 PACRG PSARL RBM28 FDPS	10
Unmatched gene hits		
Nkx2-5	CCND2 CNOT2 MARS VSIG2 DDX23	5
GATA-4	CCND2 CNOT2 MARS VSIG2 DDX23	5
CDX	CCND2 CNOT2 MARS VSIG2 DDX23	5
TFIIA	CCND2 CNOT2 MARS VSIG2 DDX23	5
HNF-3alpha	CCND2 CNOT2 MARS VSIG2 DDX23	5
Evi-1	CCND2 CNOT2 MARS VSIG2 DDX23	5
HNF-1	CCND2 CNOT2 MARS VSIG2 DDX23	5
Oct-1	CCND2 CNOT2 MARS VSIG2 DDX23	5
LXR, PXR, CAR, COUP, RAR	CCND2 CNOT2 MARS VSIG2 DDX23	5
TFII-I	CCND2 CNOT2 MARS VSIG2 DDX23	5

Supplementary Table 7.8 Comparison of DEFA1 and DEFA5 gene hits based on GO terms.

This table lists the common and different GO categories for DEFA1 and DEFA5 gene hits

Gene Ontology : biological process. Level: 4	DEFA1_genes	No. of genes	DEFA5_genes	No. of genes	Unadjusted pvalue	Adjusted pvalue FDR
COMMON GO categories						
immune response	INS PSMB8 CX3CL1	3	TTF AIF1 DHLAG HLA-DMA ILF2 G10P1 IFI30 CX3CL1	8	7.18E-01	1
macromolecule metabolism	DPM1 CHD1L H1F0 TMPRSS1 CHD2 NCL INS	25	CLG4A KIAA0175 DENR HER3 ATPAF1 SPC18	47	7.25E-01	1

	FARS1 MTMR5 HRMT1L1 ALDR1 PDCD9 KIAA0060 PSMB8 PIAS2 MARS PFD4 ARAF DDX23 MYST2 LSM6 FKBP12 KIAA0935 MARCH5 CASP5		SAP61 VPS11 TRIM9 PFKL NMP238 MKI67IP SAP114 RPS27L UGP2 MID1 NEC2 UBQLN1 RPP40 DHLAG RNMTL1 KIAA0091 PIM2 NAALAD1 MRPL4 PRIM1 AP2B1 PSMA7 ISOT EPM2A USP39 TOMM34 HSPA14 APOL1 TYRO10 KIAA0517 MRPS36 MRPS2 NUDT5 KIAA1667 AUH RBM3 TRIM38 MNK1 H2AFX TRX1 CBX6			
neurophysiological process	ARRB2 CPNE6	2	RDH1 YWHAG CKLFSF6 KCNMA1 SLUG	5	1	1
cell activation	INS	1	TTF DHLAG	2	1	1
regulation of cellular physiological process	CHD2 CCNI CNOT2 KIAA0065 PIAS2 MYST2 KIAA0929 MIZ1 CCND2 CASP5	10	PCCX1 ATPIF1 PRDX2 TTF YWHAG NMP238 MXD3 AIF1 DHLAG SLUG ILF2 ELE1 MBD1 CCNB SCAND2 EPM2A MNK1 PLAGL2 TRX1 CBX6 HOX1G	21	1	1
catabolism	RODH INS KIAA0060 PSMB8	4	CLG4A PFKL PSMA7 ISOT USP39 KIAA0106 NUDT5 AUH GM2A	9	1	1
regulation of metabolism	CHD2 INS CNOT2 KIAA0065 PIAS2 MYST2 KIAA0929 MIZ1	8	PCCX1 ATPIF1 TTF NMP238 MXD3 SLUG ILF2 ELE1 MBD1 SCAND2 EPM2A MNK1 PLAGL2 TRX1 CBX6 HOX1G	16	1	1
regulation of signal transduction	FKBP12	1	TTF YWHAG TRIM38 PAK1IP1	4	1	1
cellular metabolism	DPM1 CHD1L H1F0 TMPRSS1 CHD2 NCL RODH INS FARS1 CNOT2 MTMR5 HRMT1L1 KIAA0065 NDUFS5 PDCD9 KIAA0060 PSMB8 PIAS2 GUK1 MARS PFD4 ARAF DDX23 PECI MYST2 LSM6 FKBP12 KIAA0929 KIAA0935 MARCH5 MIZ1	33	PCCX1 CLG4A KIAA0175 ATPIF1 PRDX2 TXNRD1 DENR TTF HER3 LTB4DH TSSC3 ATPAF1 SPC18 SAP61 VPS11 TRIM9 PFKL NMP238 MKI67IP SAP114 RPS27L UGP2 MID1 MXD3 NEC2 UBQLN1 RPP40 DHLAG	67	1	1

	CASP5 KBL		RNMTL1 KIAA0091 PIM2 NAALAD1 MRPL4 PRIM1 AP2B1 SLUG ILF2 EFCBP1 ELE1 MBD1 PSMA7 SCAND2 ISOT EPM2A USP39 TOMM34 HSPA14 CEV14 KIAA0106 APOL1 TYRO10 KIAA0517 NPD002 MRPS36 MRPS2 NUDT5 KIAA1667 AUH RBM3 TRIM38 MNK1 PLAGL2 H2AFX GM2A TRX1 CBX6 HOX1G			
nervous system development	CPNE6	1	KIAA1745 YWHAG SNAI1	3	1	1
negative regulation of physiological process	INS MIZ1	2	ATPIF1 YWHAG AIF1 DHLAG SLUG MBD1	6	1	1
biosynthesis	DPM1 RODH INS FARS1 PDCD9 GUK1 MARS KBL	8	DENR RPS27L DHLAG MRPL4 EFCBP1 EPM2A MRPS36 MRPS2 KIAA1667 MNK1	10	2.88E-01	1
establishment of localization	NTT73 SEC5L1 INS BGP1 NDUFS5 ATP5S FTL CPNE6	8	TXNRD1 TTF MSTP028 VPS11 CTRP6 DHLAG KCNMA1 COL9A3 STX6 AP2B1 TFIP11 STAU2 TOMM34 APOL1 SATT BAP29 NPD002 SLC17A5 KIAA1667 VIM ARFL3 NESH	22	5.16E-01	1
protein localization	SEC5L1 INS	2	TTF VPS11 DHLAG STX6 AP2B1 TOMM34 BAP29 KIAA1667 ARFL3	9	5.05E-01	1
cell cycle	CCNI DCTN3 MIZ1 CCND2	4	YWHAG AIF1 PIM2 CCNB H2AFX	5	4.74E-01	1
primary metabolism	DPM1 CHD1L H1F0 TMPRSS1 CHD2 NCL RODH INS FARS1 CNOT2 MTMR5 HRMT1L1 KIAA0065 ALDR1 PDCD9 KIAA0060 PSMB8 PIAS2 GUK1 MARS PFD4 ARAF DDX23 PECI MYST2 LSM6 FKBP12 KIAA0929 KIAA0935 CPNE6 MARCH5 MIZ1 CASP5 KBL	34	PCCX1 CLG4A KIAA0175 ATPIF1 DENR TTF HER3 LTB4DH TSSC3 ATPAF1 SPC18 SAP61 VPS11 TRIM9 PFKL NMP238 MKI67IP SAP114 RPS27L UGP2 MID1 MXD3 NEC2 UBQLN1 RPP40 DHLAG RNMTL1 KIAA0091 PIM2 NAALAD1 MRPL4 PRIM1 AP2B1 SLUG	63	4.59E-01	1

			ILF2 ELE1 MBD1 PSMA7 SCAND2 ISOT EPM2A USP39 TOMM34 HSPA14 CEV14 KIAA0106 APOL1 TYRO10 KIAA0517 MRPS36 MRPS2 NUDT5 KIAA1667 AUH RBM3 TRIM38 MNK1 PLAGL2 H2AFX GM2A TRX1 CBX6 HOX1G			
cell organization and biogenesis	H1F0 CHD2 MYST2 PEX11G	4	TIN2 YWHAG TMSB10 NMP238 MID1 DHLAG STX6 AP2B1 TOMM34 BAP29 KIAA1667 ARFL3 H2AFX CBX6	14	4.24E-01	1
transport	NTT73 SEC5L1 INS NDUFS5 ATP5S FTL CPNE6	7	TXNRD1 TTF MSTP028 VPS11 CTRP6 DHLAG KCNMA1 COL9A3 STX6 AP2B1 TFIP11 STAU2 TOMM34 APOL1 SATT BAP29 NPD002 SLC17A5 KIAA1667 VIM ARFL3	21	3.79E-01	1
signal transduction	APBB1IP INS BGP1 HRMT1L1 FMOD PIAS2 ARAF ARRB2 FKBP12 KIAA0929	10	PDZK2 TXNRD1 TTF HER3 YWHAG DHLAG TENC1 ELE1 TYRO10 DRG2 ARFL3 TRIM38 MNK1 PAK1IP1	14	3.44E-01	1
nitrogen compound metabolism	INS FARS1 KIAA0060 MARS KBL	5	AUH	1	1.44E-02	1
cell death	INS PDCD9 CASP5	3	PRDX2 TSSC3 YWHAG DHLAG BAP29	5	7.16E-01	1
positive regulation of physiological process	INS PIAS2	2	ILF2 ELE1	2	5.97E-01	1
cell-cell signaling	INS CPNE6	2	YWHAG NEC2 KCNMA1	3	6.62E-01	1
positive regulation of physiological process	INS PIAS2	2	ILF2 ELE1	2	5.97E-01	1

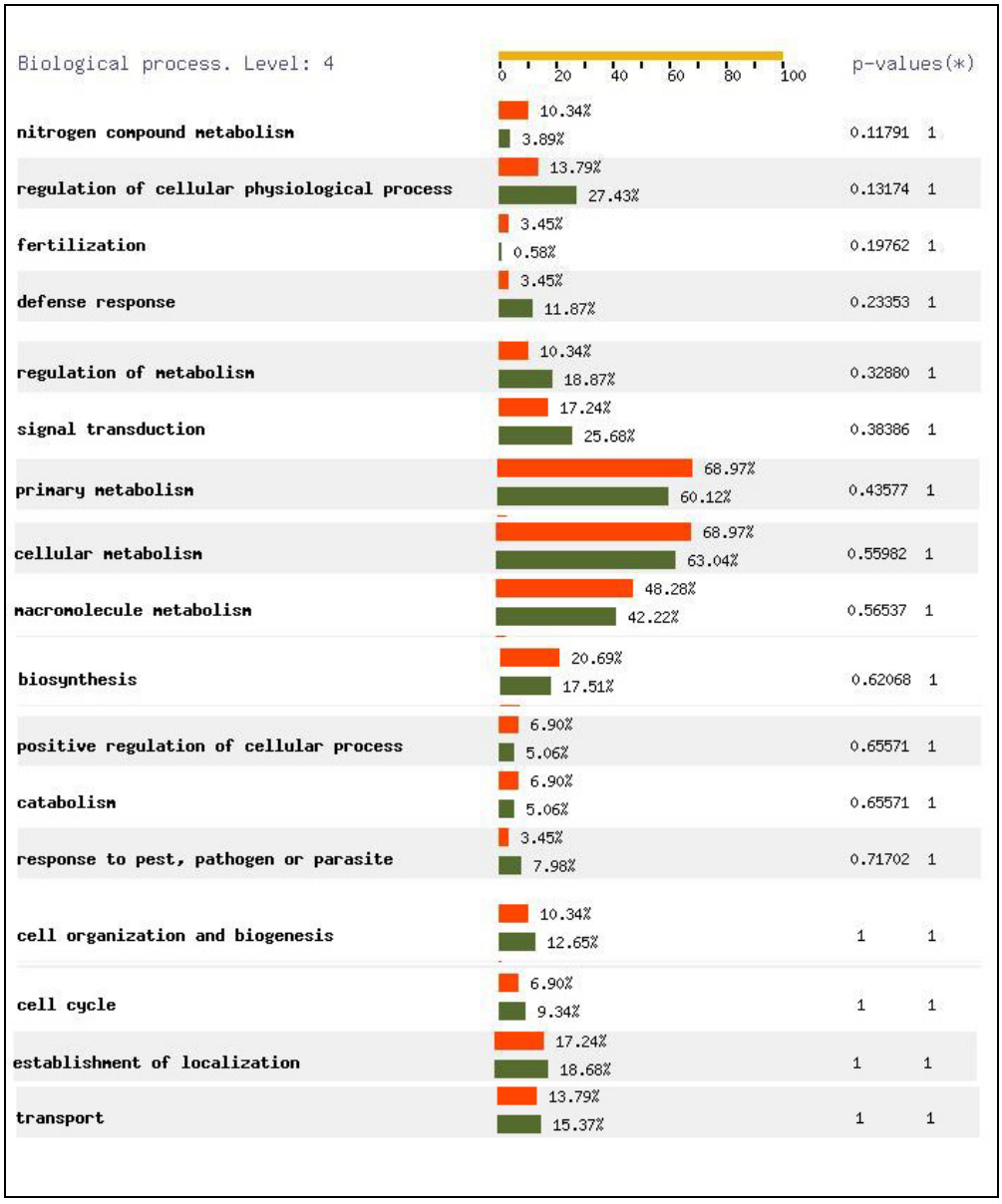
GO categories exclusive for one AMP gene group

Gene Ontology : biological process. Level: 4	DEFA1_genes	No. of genes	DEFA5_genes	No. of genes	Unadjusted pvalue	Adjusted pvalue FDR
cellular localization	No genes	0	DHLAG STX6 AP2B1 TOMM34 BAP29 KIAA1667 ARFL3	7	9.62E-02	1
reproductive organismal physiological process	BGP1	1	No genes	0	3.27E-01	1

vasculature development	BGP1	1	No genes	0	3.27E-01	1
ion homeostasis	FTL	1	No genes	0	3.27E-01	1
ectoderm development	KRT5	1	No genes	0	3.27E-01	1
cell proliferation	No genes	0	AIF1 DHLAG PIM2	3	5.51E-01	1
hemopoietic or lymphoid organ development	No genes	0	TTF DHLAG TRX1	3	5.51E-01	1
cartilage development	No genes	0	SNAI1	1	1	1
skeletal development	No genes	0	SNAI1 TFIP11	2	1	1
male sex differentiation	No genes	0	ELE1	1	1	1
lymphocyte differentiation	No genes	0	TTF DHLAG	2	1	1
cellular morphogenesis	No genes	0	NMP238	1	1	1
regulation of cell growth	No genes	0	NMP238	1	1	1
regulation of developmental pigmentation	No genes	0	KIAA1667	1	1	1
neuron differentiation	No genes	0	YWHAG	1	1	1
response to oxidative stress	No genes	0	PRDX2 KIAA0106	2	1	1
response to wounding	No genes	0	AIF1 DHLAG	2	1	1
embryonic hemopoiesis	No genes	0	TRX1	1	1	1
cell growth	No genes	0	NMP238	1	1	1
positive regulation of development	No genes	0	KIAA1667	1	1	1
mesoderm development	No genes	0	SLUG	1	1	1
development of primary sexual characteristics	No genes	0	ELE1	1	1	1
embryonic organ development	No genes	0	TRX1	1	1	1
taxis	No genes	0	CKLFSF6	1	1	1
regulation of cell differentiation	No genes	0	YWHAG	1	1	1
regulation of hydrolase activity	No genes	0	HRBL	1	1	1

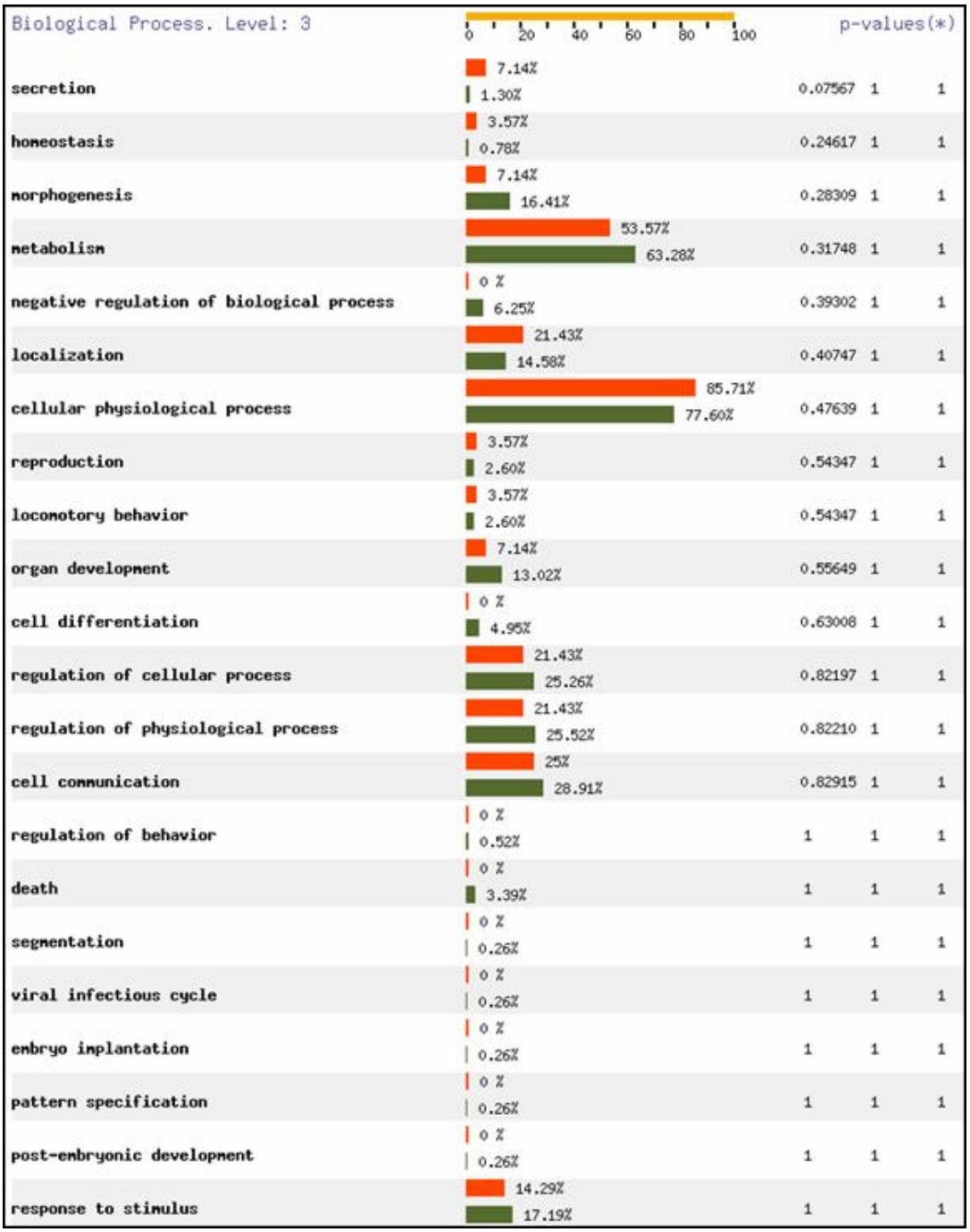
Figures for Chapter 7

Supplementary Figure 7.1: Alpha defensin 1 unmatched gene hits (did not match with co-expressed gene list for DEFA1, DEFA3) compared with co-expressed genes of DEFA1,DEFA3



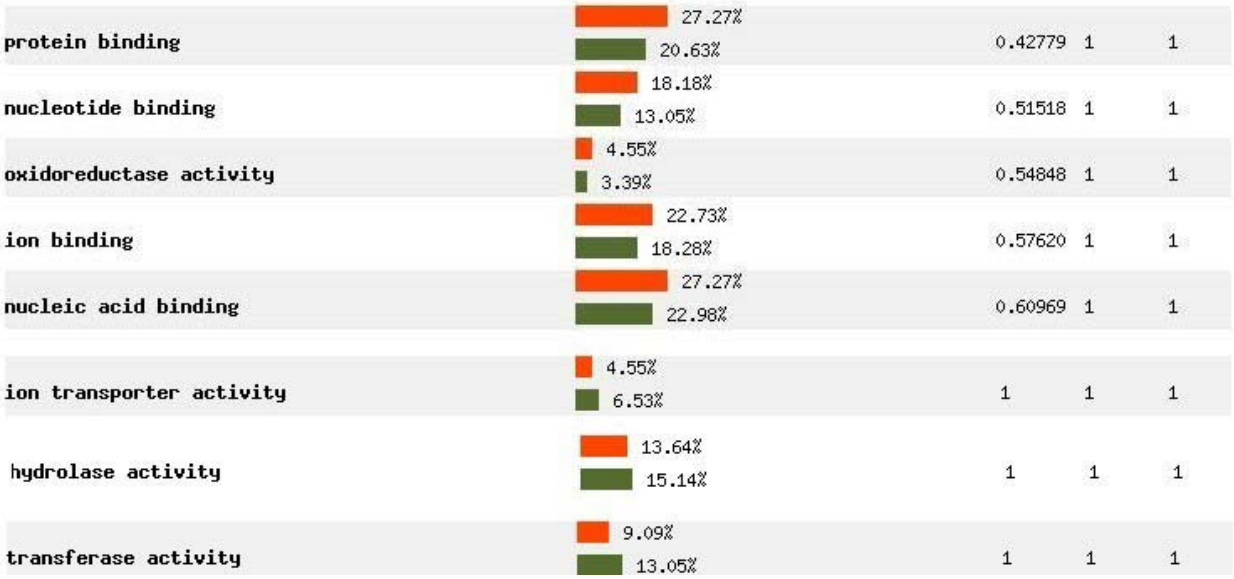
Supplementary Figure 7.1: The orange bar indicates the unmatched gene hits and the green bar indicates the co-expressed genes for DEFA1, DEFA3. The raw p-values and corrected p-values are shown in column 3 and column 4 respectively

Supplementary Figure 7.2: All alpha defensin 1 predicted genes compared with co-expressed genes in terms of GO biological function



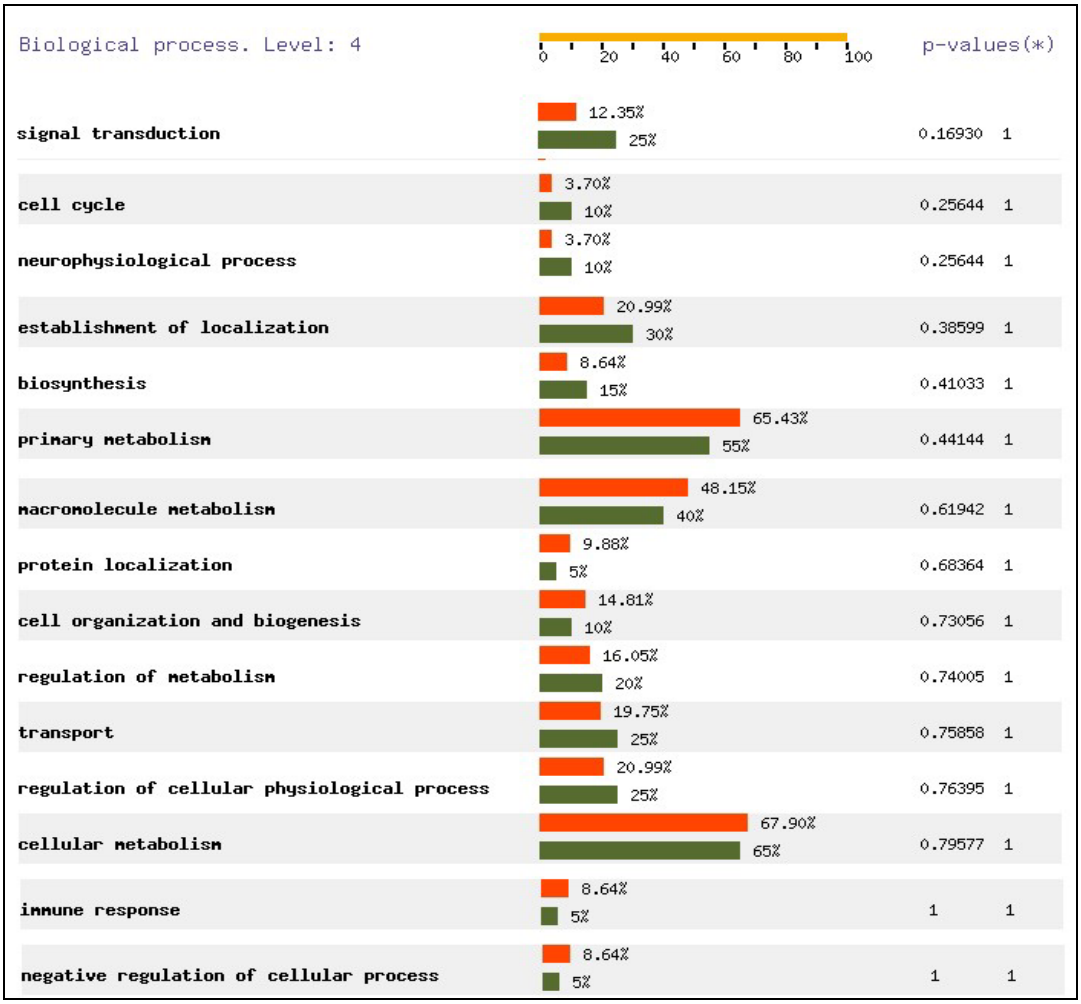
Supplementary Figure 7.2: The orange bar represents the co-expressed genes and green bar represents predicted genes

Supplementary Figure 7.3: All alpha defensin 1 predicted genes compared with co-expressed genes in terms of GO molecular function



Supplementary Figure 7.3:. The orange bar indicates the co-expressed genes and green bar indicates predictedgenes

Supplementary Figure 7.4: DEFA4 novel predicted genes compared with matched predicted genes grouped based on GO biological function



Supplementary Figure 7.4: The orange bar is the matched predicted genes and green bar is unmatched predicted genes

Supplementary References

Supplementary References for Table 5.3 and Table 6.1

Albig, W., Trappe, R., *et al.*,1999. The human H2A and H2B histone gene complement. Biol Chem 380, 7-18.

Courselaud, B., Pigeon, C., *et al.*,2002. C/EBPalpha regulates hepatic transcription of hepcidin, an antimicrobial peptide and regulator of iron metabolism. Cross-talk between C/EBP pathway and iron metabolism. J Biol Chem 277, 41163-41170.

Deen, P. M., Terwel, D., *et al.*,1991. Structural analysis of the entire proopiomelanocortin gene of Xenopus laevis. Eur J Biochem 201, 129-137.

Elholm, M., Bjerking, G., *et al.*,1996. Regulatory elements in the promoter region of the rat gene encoding the acyl-CoA-binding protein. Gene 173, 233-238.

Frohm Nilsson, M., Sandstedt, B., *et al.*,1999. The human cationic antimicrobial protein (hCAP18), a peptide antibiotic, is widely expressed in human squamous epithelia and colocalizes with interleukin-6. Infect Immun 67, 2561-2566.

Fu, W., Shah, S. R., *et al.*,1997. Transactivation of proenkephalin gene by HTLV-1 tax1 protein in glial cells: involvement of Fos/Jun complex at an AP-1 element in the proenkephalin gene promoter. J Neurovirol 3, 16-27.

Hahm, S. H. and Eiden, L. E.,1998. Cis-regulatory elements controlling basal and inducible VIP gene transcription. Ann N Y Acad Sci 865, 10-26.

Harder, J., Meyer-Hoffert, U., *et al.*,2000. Mucoid Pseudomonas aeruginosa, TNF-alpha, and IL-1beta, but not IL-6, induce human beta-defensin-2 in respiratory epithelia. Am J Respir Cell Mol Biol 22, 714-721.

Hayashi, R., Wada, H., *et al.*,2004. Effects of glucocorticoids on gene transcription. Eur J Pharmacol 500, 51-62.

Hocker, M., Raychowdhury, R., *et al.*,1998. Sp1 and CREB mediate gastrin-dependent regulation of chromogranin A promoter activity in gastric carcinoma cells. J Biol Chem 273, 34000-34007.

Huang, C. J., Nazarian, R., *et al.*,2002. Tumor necrosis factor modulates transcription of myelin basic protein gene through nuclear factor kappa B in a human oligodendroglioma cell line. Int J Dev Neurosci 20, 289-296.

Khanna-Gupta, A., Zibello, T., *et al.*,2000. Sp1 and C/EBP are necessary to activate the lactoferrin gene promoter during myeloid differentiation. Blood 95, 3734-3741.

King, A. E., Morgan, K., *et al.*,2003. Differential regulation of secretory leukocyte protease inhibitor and elafin by progesterone. Biochem Biophys Res Commun 310, 594-599.

Kobierski, L. A., Wong, A. E., *et al.*,1999. Cyclic AMP-dependent activation of the proenkephalin gene requires phosphorylation of CREB at serine-133 and a Src-related kinase. J Neurochem 73, 129-138.

Le, Y., Gagneten, S., *et al.*,2003. Far-upstream elements are dispensable for tissue-specific proenkephalin expression using a Cre-mediated knock-in strategy. J Neurochem 84, 689-697.

Lennartsson, A., Pieters, K., *et al.*,2003. AML-1, PU.1, and Sp3 regulate expression of human bactericidal/permeability-increasing protein. Biochem Biophys Res Commun 311, 853-863.

Liu, F., Kondova, I., *et al.*,2000. Detection of PACH1, a nuclear factor implicated in the transcriptional regulation of meiotic and early haploid stages of spermatogenesis. Mol Reprod Dev 57, 224-231.

Lu, Z., Kim, K. A., *et al.*,2004. MEF up-regulates human beta-defensin 2 expression in epithelial cells. FEBS Lett 561, 117-121.

Mahapatra, N. R., Mahata, M., *et al.*,2003. Secretin activation of chromogranin A gene transcription. Identification of the signaling pathways in cis and in trans. J Biol Chem 278, 19986-19994.

Mahata, S. K., Mahapatra, N. R., *et al.*,2002. Neuroendocrine cell type-specific and inducible expression of chromogranin/secretogranin genes: crucial promoter motifs. Ann N Y Acad Sci 971, 27-38.

Nguyen, H., Teskey, L., *et al.*,1999. Identification of the secretory leukocyte protease inhibitor (SLPI) as a target of IRF-1 regulation. Oncogene 18, 5455-5463.

Oswald, F., Dobner, T., *et al.*,1996. The E2F transcription factor activates a replication-dependent human H2A gene in early S phase of the cell cycle. Mol Cell Biol 16, 1889-1895.

Persson, P., Manetopoulos, C., *et al.*,2004. Olf/EBF proteins are expressed in neuroblastoma cells: potential regulators of the Chromogranin A and SCG10 promoters. Int J Cancer 110, 22-30.

Pohl, T. M., Phillips, E., *et al.*,1990. The organisation of the mouse chromogranin B (secretogranin I) gene. FEBS Lett 262, 219-224.

Rozansky, D. J., Wu, H., *et al.*,1994. Glucocorticoid activation of chromogranin A gene expression. Identification and characterization of a novel glucocorticoid response element. J Clin Invest 94, 2357-2368.

Sandberg, M. B., Bloksgaard, M., *et al.*,2005. The gene encoding acyl-CoA-binding protein is subject to metabolic regulation by both sterol regulatory element-binding protein and peroxisome proliferator-activated receptor alpha in hepatocytes. J Biol Chem 280, 5258-5266.

Slutsky, S. G., Kamaraju, A. K., *et al.*,2003. Activation of myelin genes during transdifferentiation from melanoma to glial cell phenotype. J Biol Chem 278, 8960-8968.

Suico, M. A., Koga, T., *et al.*,2004. Sp1 is involved in the transcriptional activation of lysozyme in epithelial cells. Biochem Biophys Res Commun 324, 1302-1308.

Teng, C. T.,2002. Lactoferrin gene expression and regulation: an overview. Biochem Cell Biol 80, 7-16.

Trappe, R., Doenecke, D., *et al.*,1999. The expression of human H2A-H2B histone gene pairs is regulated by multiple sequence elements in their joint promoters. Biochim Biophys Acta 1446, 341-351.

Vora, P., Youdim, A., *et al.*,2004. Beta-defensin-2 expression is regulated by TLR signaling in intestinal epithelial cells. J Immunol 173, 5398-5405.

Wang, T. T., Nestel, F. P., *et al.*,2004. Cutting edge: 1,25-dihydroxyvitamin D3 is a direct inducer of antimicrobial peptide gene expression. J Immunol 173, 2909-2912.

Wei, Q., Miskimins, W. K., *et al.*,2003. Cloning and characterization of the rat myelin basic protein gene promoter. Gene 313, 161-167.

Wei, Q., Miskimins, W. K., *et al.*,2004. Sox10 acts as a tissue-specific transcription factor enhancing activation of the myelin basic protein gene promoter by p27Kip1 and Sp1. J Neurosci Res 78, 796-802.

Wei, Q., Miskimins, W. K., *et al.*,2005. Stage-specific expression of myelin basic protein in oligodendrocytes involves Nkx2.2-mediated repression that is relieved by the Sp1 transcription factor. J Biol Chem 280, 16284-16294.

Wu, H., Zhang, G., *et al.*,2000. Regulation of cathelicidin gene expression: induction by lipopolysaccharide, interleukin-6, retinoic acid, and Salmonella enterica serovar typhimurium infection. Infect Immun 68, 5552-5558.

Yamamoto, C. M., Banaiee, N., *et al.*,2004. Alpha-defensin expression during myelopoiesis: identification of cis and trans elements that regulate expression of NP-3 in rat promyelocytes. J Leukoc Biol 75, 332-341.

Appendices

Great spirits have always encountered violent opposition from mediocre minds.
(Albert Einstein)

Appendix 1

Supplementary Material for Chapter 4

Figure 4.1: Melittin profile query profile results:

```
Query sequence: mellitin1
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                       Score    E-value  N
-------- -----------                                       -----    ------- ---
mellitin                                                    63.9    5.8e-20   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
mellitin   1/1       1    26 [.     1    26 []    63.9  5.8e-20

Alignments of top-scoring domains:
mellitin: domain 1 of 1, from 1 to 26: score 63.9, E = 5.8e-20
                     *->GiGAiLKvLatGLPaLisWiKrKRqq<-*
                        GiGA+LKVL+tGLPaLisWiKrKRqq
   mellitin1     1    GIGAVLKVLTTGLPALISWIKRKRQQ    26

//

Query sequence: mut5_L6
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                       Score    E-value  N
-------- -----------                                       -----    ------- ---
mellitin                                                    59.5    1.2e-18   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
mellitin   1/1       1    26 []     1    26 []    59.5  1.2e-18

Alignments of top-scoring domains:
mellitin: domain 1 of 1, from 1 to 26: score 59.5, E = 1.2e-18
                     *->GiGAiLKvLatGLPaLisWiKrKRqq<-*
                        GiGA+ KVL+tGLPaLisWiKrKRqq
     mut5_L6     1    GIGAVWKVLTTGLPALISWIKRKRQQ    26
```

```
Query sequence: mut13_L13
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                       Score    E-value  N
-------- -----------                                       -----    ------- ---
mellitin                                                    59.5    1.2e-18   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
mellitin   1/1       1    26 []     1    26 []    59.5  1.2e-18

Alignments of top-scoring domains:
mellitin: domain 1 of 1, from 1 to 26: score 59.5, E = 1.2e-18
                     *->GiGAiLKvLatGLPaLisWiKrKRqq<-*
                        GiGA+LKvL+tG PaLisWiKrKRqq
   mut13_L13     1    GIGAVLKVLTTGWPALISWIKRKRQQ    26

//

Query sequence: mut1_G1
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                       Score    E-value  N
-------- -----------                                       -----    ------- ---
mellitin                                                    59.1    1.6e-18   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
mellitin   1/1       2    26 .]     1    26 []    59.1  1.6e-18

Alignments of top-scoring domains:
mellitin: domain 1 of 1, from 2 to 26: score 59.1, E = 1.6e-18
                     *->GiGAiLKvLatGLPaLisWiKrKRqq<-*
                         iGA+LKvL+tGLPaLisWiKrKRqq
     mut1_G1     2    -IGAVLKVLTTGLPALISWIKRKRQQ    26
```

```
Query sequence: mut6_L7
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                        Score    E-value  N
-------- -----------                                        -----    ------- ---
mellitin                                                     58.6    2.3e-18   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
mellitin   1/1       1    26 []     1    26 []    58.6  2.3e-18

Alignments of top-scoring domains:
mellitin: domain 1 of 1, from 1 to 26: score 58.6, E = 2.3e-18
                     *->GiGAiLKvLatGLPaLisWiKrKRqq<-*
                        GiGA+L vL+tGLPaLisWiKrKRqq
     mut6_L7     1    GIGAVLWVLTTGLPALISWIKRKRQQ    26

//

Query sequence: mut10_T11
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                        Score    E-value  N
-------- -----------                                        -----    ------- ---
mellitin                                                     58.1    3.4e-18   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
mellitin   1/1       1    26 []     1    26 []    58.1  3.4e-18

Alignments of top-scoring domains:
mellitin: domain 1 of 1, from 1 to 26: score 58.1, E = 3.4e-18
                     *->GiGAiLKvLatGLPaLisWiKrKRqq<-*
                        GiGA+LKvL+ GLPaLisWiKrKRqq
   mut10_T11     1    GIGAVLKVLTWGLPALISWIKRKRQQ    26
```

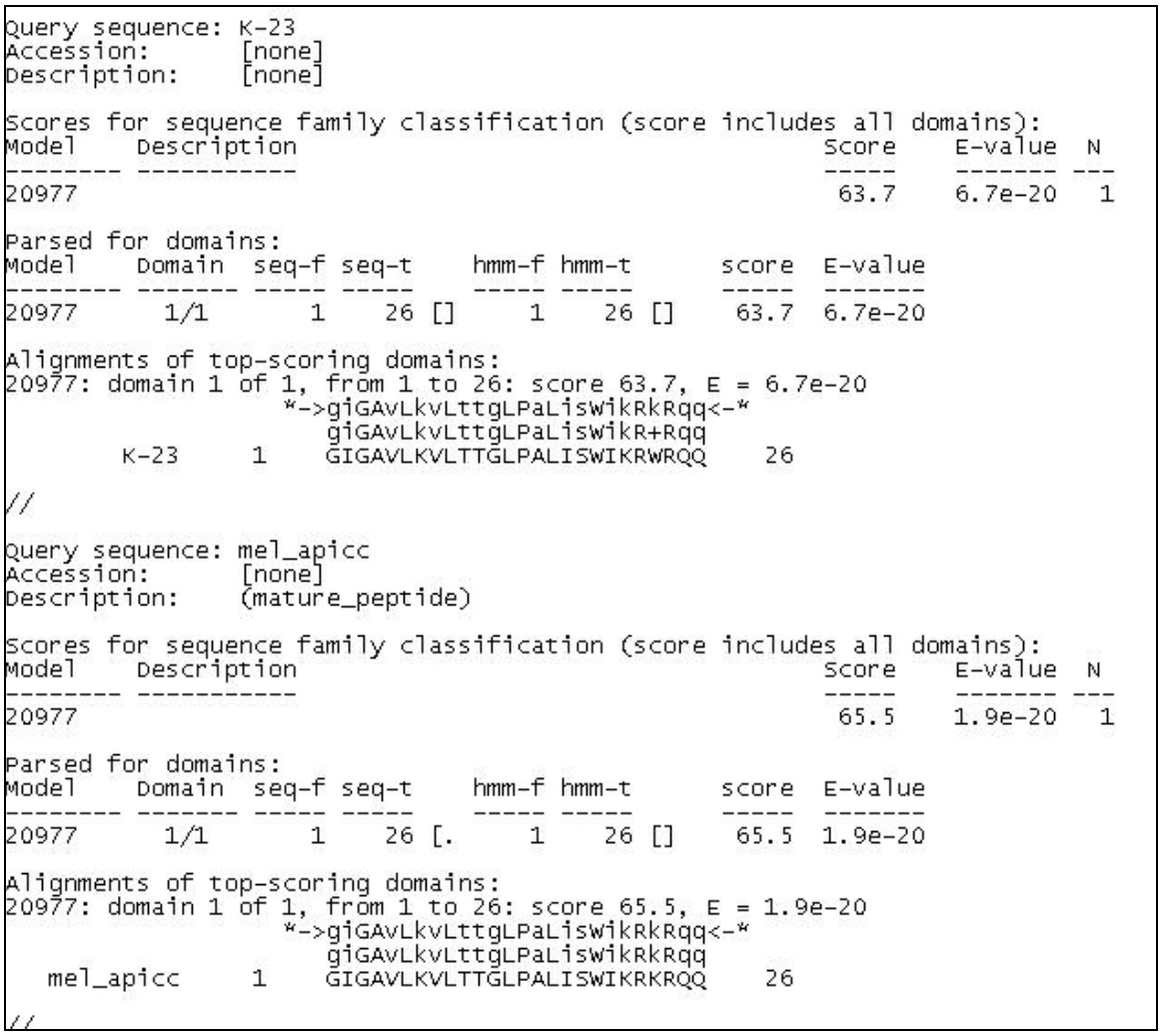
```
Query sequence: mut13_P14
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                        Score    E-value  N
-------- -----------                                        -----    ------- ---
mellitin                                                     56.9    7.2e-18   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
mellitin   1/1       1    26 []     1    26 []    56.9  7.2e-18

Alignments of top-scoring domains:
mellitin: domain 1 of 1, from 1 to 26: score 56.9, E = 7.2e-18
                     *->GiGAiLKvLatGLPaLisWiKrKRqq<-*
                        GiGA+LKvL+tGL aLisWiKrKRqq
   mut13_P14     1    GIGAVLKVLTTGLWALISWIKRKRQQ    26

//

Query sequence: Cecropin
Accession:      [none]
Description:    A (1-8)-Melittin (1-18)

Scores for sequence family classification (score includes all domains):
Model    Description                                        Score    E-value  N
-------- -----------                                        -----    ------- ---
mellitin                                                     20.9    5.2e-07   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
mellitin   1/1       9    26 .]     1    26 []    20.9  5.2e-07

Alignments of top-scoring domains:
mellitin: domain 1 of 1, from 9 to 26: score 20.9, E = 5.2e-07
                     *->GiGAiLKvLatGLPaLisWiKrKRqq<-*
                        GiGA+LKvL+tGLPaLis
    Cecropin     9    GIGAVLKVLTTGLPALIS--------    26
```

```
Query sequence: CA(1-7)M(2-9)
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                        Score    E-value  N
-------- -----------                                        -----    ------- ---
mellitin                                                    -19.1          1   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
mellitin   1/1       8    15 .]     1    26 []   -19.1        1

Alignments of top-scoring domains:
mellitin: domain 1 of 1, from 8 to 15: score -19.1, E = 1
                     *->GiGAiLKVLatGLPaLisWiKrKRqq<-*
                         iGA+LKVL
  CA(1-7)M(2     8    -IGAVLKVL-----------------    15

//

Query sequence: Protegrin
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                        Score    E-value  N
-------- -----------                                        -----    ------- ---
mellitin                                                    -33.7          1   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
mellitin   1/1       1    12 [.     1    26 []   -33.7        1

Alignments of top-scoring domains:
mellitin: domain 1 of 1, from 1 to 12: score -33.7, E = 1
                     *->GiGAiLKVLatGLPaLisWiKrKRqq<-*
                                GL     + +r+
   Protegrin     1    ----------GGL----CYCRRRFCV    12
```

```
Query sequence: Acyl-CoA
Accession:      [none]
Description:    dehydrogenase family member 8

Scores for sequence family classification (score includes all domains):
Model    Description                                        Score    E-value  N
-------- -----------                                        -----    ------- ---
mellitin                                                     -6.0      0.052   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
mellitin   1/1     214   236 ..     1    26 []    -6.0    0.052

Alignments of top-scoring domains:
mellitin: domain 1 of 1, from 214 to 236: score -6.0, E = 0.052
                    *->GiGAiLKvLatGLPaLisWiKrKRqq<-*
                       Gi  i  v+  G P L s+ K+ ++
    Acyl-CoA   214    GISCI--VVEKGTPGL-SFGKKEKKV    236

//

Query sequence: mellitin_complete
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                        Score    E-value  N
-------- -----------                                        -----    ------- ---
mellitin                                                    63.9    5.8e-20   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
mellitin   1/1      44    69 ..     1    26 []    63.9  5.8e-20

Alignments of top-scoring domains:
mellitin: domain 1 of 1, from 44 to 69: score 63.9, E = 5.8e-20
                    *->GiGAiLKvLatGLPaLisWiKrKRqq<-*
                       GiGA+LKvL+tGLPaLisWiKrKRqq
  mellitin_c    44    GIGAVLKVLTTGLPALISWIKRKRQQ    69
```

Figure 4.1: The mellitin profile is tested against a set of 12 sequences which include mel_apicc (mature_peptide), melittin analogs: mut5_l6, mut13_l13, mut1_g1, mut6_l7, mut10_t11, mut13_p14, melittin hybrid: cecropina(1-8)-melittin(1-18), ca(1-7)m(2-9), non-melittin sequences: protegrin (PG3_PIG), acyl-coadehydrogenasefamilymember8 (ACAD8_HUMAN), mel_apicc(complete peptide) (melittin_complete). The E-value and score indicate the statistical significance of similarity of the sequence to the profile. A lower E-value score indicates a better match. Analysis of the E-values of different test sequences shows that the melittin profile generated by HMM is able to differentiate between members of the melittin family and non–members.

Figure 4.2: Melittin analog profile analysis



```
Query sequence: K-23
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                        Score    E-value  N
-------- -----------                                        -----    ------- ---
20977                                                        63.7    6.7e-20   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
20977      1/1       1    26 []     1    26 []    63.7  6.7e-20

Alignments of top-scoring domains:
20977: domain 1 of 1, from 1 to 26: score 63.7, E = 6.7e-20
                     *->giGAvLkvLttgLPaLisWikRkRqq<-*
                        giGAvLkvLttgLPaLisWikR+Rqq
        K-23     1    GIGAVLKVLTTGLPALISWIKRWRQQ    26

//

Query sequence: mel_apicc
Accession:      [none]
Description:    (mature_peptide)

Scores for sequence family classification (score includes all domains):
Model    Description                                        Score    E-value  N
-------- -----------                                        -----    ------- ---
20977                                                        65.5    1.9e-20   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
20977      1/1       1    26 [.     1    26 []    65.5  1.9e-20

Alignments of top-scoring domains:
20977: domain 1 of 1, from 1 to 26: score 65.5, E = 1.9e-20
                     *->giGAvLkvLttgLPaLisWikRkRqq<-*
                        giGAvLkvLttgLPaLisWikRkRqq
   mel_apicc     1    GIGAVLKVLTTGLPALISWIKRKRQQ    26

//
```

```
Query sequence: L-16
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                       Score    E-value  N
-------- -----------                                       -----    ------- ---
20977                                                       60.7    5.2e-19   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
20977      1/1       1    26 []     1    26 []    60.7  5.2e-19

Alignments of top-scoring domains:
20977: domain 1 of 1, from 1 to 26: score 60.7, E = 5.2e-19
                     *->giGAvLkvLttgLPaLisWikRkRqq<-*
                        giGAvLkvLttgLPa isWikRkRqq
        L-16     1    GIGAVLKVLTTGLPAWISWIKRKRQQ    26

//

Query sequence: I-2
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                       Score    E-value  N
-------- -----------                                       -----    ------- ---
20977                                                       60.0    8.4e-19   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
20977      1/1       1    26 []     1    26 []    60.0  8.4e-19

Alignments of top-scoring domains:
20977: domain 1 of 1, from 1 to 26: score 60.0, E = 8.4e-19
                     *->giGAvLkvLttgLPaLisWikRkRqq<-*
                        g GAvLkvLttgLPaLisWikRkRqq
         I-2     1    GWGAVLKVLTTGLPALISWIKRKRQQ    26
```

Figure 4.2: The mellitin profiles categorizing decreased hemolytic activity, increased hemolytic activity is tested against a set of melittin analogs, K-23, L-16, I-2 and normal melittin sequence melittin wild type, mel_apicc (mature_peptide). The profiles could distinguish between mutants with decreased hemolytic activity and increased hemolytic activity.

Figure 4.3: Beta-defensin profile query profile results

```
Query sequence: BD01_CERPR
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                       Score    E-value  N
-------- -----------                                       -----    ------- ---
22494                                                      102.5    1.4e-31   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
22494      1/1       1    36 []     1    36 []   102.5  1.4e-31

Alignments of top-scoring domains:
22494: domain 1 of 1, from 1 to 36: score 102.5, E = 1.4e-31
                     *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                        DhY+Cv+sGGqCLySaCPiyTKiQGTCy+gKakCCK
  BD01_CERPR     1    DHYNCVRSGGQCLYSACPIYTKIQGTCYHGKAKCCK    36
```

```
Query sequence: BD01_PONPY
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                       Score    E-value  N
-------- -----------                                       -----    ------- ---
22494                                                      101.1    3.7e-31   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
22494      1/1       2    37 .]     1    36 []   101.1  3.7e-31

Alignments of top-scoring domains:
22494: domain 1 of 1, from 2 to 37: score 101.1, E = 3.7e-31
                     *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                        DhY+Cv+sGGqCLySaCPi+TKiQGTCy+gKakCCK
  BD01_PONPY     2    DHYNCVSSGGQCLYSACPIFTKIQGTCYRGKAKCCK    37
```

```
Query sequence: BD01_CAPHI
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                       Score    E-value  N
-------- -----------                                       -----    ------- ---
22494                                                       10.1    1.7e-06   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
22494      1/1       5    40 ..     1    36 []    10.1  1.7e-06

Alignments of top-scoring domains:
22494: domain 1 of 1, from 5 to 40: score 10.1, E = 1.7e-06
                     *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                          +  C +  G C    CP + +  GTC     kCC+
  BD01_CAPHI     5    SRRSCHRNKGVCALTRCPRNMRQIGTCFGPPVKCCR    40
```

```
Query sequence: Acyl-CoA
Accession:      [none]
Description:    dehydrogenase family member 8

Scores for sequence family classification (score includes all domains):
Model    Description                                       Score    E-value  N
-------- -----------                                       -----    ------- ---
22494                                                      -28.6       0.37   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
22494      1/1     181   218 ..     1    36 []   -28.6     0.37

Alignments of top-scoring domains:
22494: domain 1 of 1, from 181 to 218: score -28.6, E = 0.37
                     *->DhYkCvqsGGqCLySaCP...iyTKiQGTCypg.KakCCK<-*
                        DhY  +  G     S   +++iy  +  T  pg+K   C
    Acyl-CoA   181    DHY--ILNGSKAFISGAGesdIYVVMCRTGGPGpKGISCI    218

//

Query sequence: Protegrin
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model    Description                                       Score    E-value  N
-------- -----------                                       -----    ------- ---
22494                                                      -40.7          1   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
22494      1/1       2    13 ..     1    36 []   -40.7        1

Alignments of top-scoring domains:
22494: domain 1 of 1, from 2 to 13: score -40.7, E = 1
                     *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                                              G Cy  +  C
   Protegrin     2    ------------------------GLCYCRRRFCVC    13
```

Figure 4.3: The beta-defensin profile is tested against a set of five sequences co-adehydrogenase family member 8, Protegrin, bd01_cerpr, bd01_caphi, bd01_ponpy . Analysis of the E-values of different test sequences shows that the beta-defensin profile generated by HMM is able to differentiate between members of the beta-defensin family and non–members.

Figure 4.4: Melittin query db results

```
HMM file:                 ./tmp/mellitin.hmm [mellitin]
Sequence database:        ./tmp/nr.db
per-sequence score cutoff:  [none]
per-domain score cutoff:    [none]
per-sequence Eval cutoff:   <= 10
per-domain Eval cutoff:     [none]
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query HMM:   mellitin
Accession:   [none]
Description: [none]
  [HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):
Sequence                              Description       Score    E-value  N
--------                              -----------       -----    ------- ---
gi|69550|pir||MEHBCI                  melittin, major    63.9      8e-15   1
gi|229444|prf||730527A                melittin,pro       51.0    6.1e-11   1
gi|16121500|ref|NP_404813.1|          tyrosine-specif    12.7        7.2   1

Parsed for domains:
Sequence                              Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
--------                              ------- ----- -----    ----- -----      -----  -------
gi|69550|pir||MEHBCI                    1/1       1    26 []     1    26 []    63.9    8e-15
gi|229444|prf||730527A                  1/1       9    34 .]     1    26 []    51.0  6.1e-11
gi|16121500|ref|NP_404813.1|            1/1     338   364 ..     1    26 []    12.7      7.2

Alignments of top-scoring domains:
gi|69550|pir||MEHBCI: domain 1 of 1, from 1 to 26: score 63.9, E = 8e-15
                    *->GiGAiLKvLatGLPaLisWiKrKRqq<-*
                       GiGA+LKvL+tGLPaLisWiKrKRqq
  gi|69550|p     1    GIGAVLKVLTTGLPALISWIKRKRQQ    26

gi|229444|prf||730527A: domain 1 of 1, from 9 to 34: score 51.0, E = 6.1e-11
                    *->GiGAiLKvLatGLPaLisWiKrKRqq<-*
                       GiG +LKvL+tGLP LisWiK K qq
  gi|229444|     9    GIGXVLKVLTTGLPXLISWIKXKXQQ    34

gi|16121500|ref|NP_404813.1|: domain 1 of 1, from 338 to 364: score 12.7, E = 7.2
                    *->GiGAi.LKvLatGLPaLisWiKrKRqq<-*
                       G  A+ L vLa  LPa+++W  rK +q
  gi|1612150   338    GFAAVaLSVLALILPAMLAWKARKLHQ    364
```

Figure 4.4: The melittin profile was queried against the nr database and three sequences were extracted by the profile. Two of the sequences were melittin sequences.

Figure 4.5: Beta-defensin querydb results

```
HMM file:                 ./tmp/28140.hmm [28140]
Sequence database:        ./tmp/nr.db
per-sequence score cutoff:  [none]
per-domain score cutoff:    [none]
per-sequence Eval cutoff:   <= 10
per-domain Eval cutoff:     [none]
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query HMM:   28140
Accession:   [none]
Description: [none]
  [HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):
Sequence                          Description        Score    E-value  N
--------                          -----------        -----    ------- ---
gi|28268769|dbj|BAC56888.1|       beta-defensin-1     99.2    1.8e-25   1
gi|4826692|ref|NP_004933.1|       defensin, beta 4    11.0       0.55   1
gi|298775|gb|AAB25873.1|          beta-defensin {p     9.8       0.77   1
gi|230338|pdb|1TAB|E              Chain E, Trypsin     9.6       0.82   1
gi|1168638|sp|P46170|BD12_BOVIN   Beta-defensin 12     8.9       0.99   1
gi|298772|gb|AAB25870.1|          beta-defensin {p     7.1        1.7   1
gi|5921174|sp|P46167|BD09_BOVIN   Beta-defensin 9      6.4        2.1   1
gi|28628181|gb|AAO32801.1|        beta-defensin-1      6.0        2.3   1
gi|15826275|pdb|1E4Q|A            Chain A, Solutio     3.5        4.8   1
gi|9957108|gb|AAG09211.1|AF181951_1  Gal-1 alpha [Gal  3.0        5.5   1
gi|9789929|ref|NP_062702.1|       defensin beta 4;      2.8        5.8   1
gi|9971962|gb|AAG10514.1|AF288371_1  beta-defensin 4   2.8        5.8   1

Parsed for domains:
Sequence                          Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
--------                          ------- ----- -----    ----- -----      -----  -------
gi|28268769|dbj|BAC56888.1|         1/1      12    47 .]     1    36 []    99.2  1.8e-25
gi|4826692|ref|NP_004933.1|         1/1      27    62 ..     1    36 []    11.0     0.55
gi|298775|gb|AAB25873.1|            1/1       6    41 .]     1    36 []     9.8     0.77
gi|230338|pdb|1TAB|E                1/1     108   139 ..     1    36 []     9.6     0.82
gi|1168638|sp|P46170|BD12_BOVIN     1/1       4    36 ..     1    36 []     8.9     0.99
gi|298772|gb|AAB25870.1|            1/1       6    41 .]     1    36 []     7.1      1.7
gi|5921174|sp|P46167|BD09_BOVIN     1/1      20    55 .]     1    36 []     6.4      2.1
gi|28628181|gb|AAO32801.1|          1/1      27    62 ..     1    36 []     6.0      2.3
gi|15826275|pdb|1E4Q|A              1/1       3    35 ..     1    36 []     3.5      4.8
gi|9957108|gb|AAG09211.1|AF181951_1  1/1      28    61 ..     1    36 []     3.0      5.5
gi|9789929|ref|NP_062702.1|         1/1      27    61 ..     1    36 []     2.8      5.8
gi|9971962|gb|AAG10514.1|AF288371_1  1/1      27    61 ..     1    36 []     2.8      5.8
```

```
gi|9971962|gb|AAG10514.1|AF288371_1    1/1      27    61 ..     1    36 []     2.8      5.8

Alignments of top-scoring domains:
gi|28268769|dbj|BAC56888.1|: domain 1 of 1, from 12 to 47: score 99.2, E = 1.8e-25
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                       DhY+C++sGGqCLySaCPi+TKiQGTCy+gKakCCK
  gi|2826876    12    DHYNCISSGGQCLYSACPIFTKIQGTCYRGKAKCCK    47

gi|4826692|ref|NP_004933.1|: domain 1 of 1, from 27 to 62: score 11.0, E = 0.55
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                       D   C++sG  C    CP   K  GTC     kCCK
  gi|4826692    27    DPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCK    62

gi|298775|gb|AAB25873.1|: domain 1 of 1, from 6 to 41: score 9.8, E = 0.77
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                            C    G CL   CP   +  GTC   + kCC+
  gi|298775|     6    SYLSCWGNRGICLLNRCPGRMRQIGTCLAPRVKCCR    41

gi|230338|pdb|1TAB|E: domain 1 of 1, from 108 to 139: score 9.6, E = 0.82
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                            C + G qCL S    +TK  GT yp+  kC K
  gi|230338|   108    ---SCASAGTQCLISGW-GNTKSSGTSYPDVLKCLK    139

gi|1168638|sp|P46170|BD12_BOVIN: domain 1 of 1, from 4 to 36: score 8.9, E = 0.99
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                            C + GG C    CP+  +  GTC     kCC+
  gi|1168638     4    ---SCGRNGGVCIPIRCPVPMRQIGTCFGRPVKCCR    36

gi|298772|gb|AAB25870.1|: domain 1 of 1, from 6 to 41: score 7.1, E = 1.7
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                            C    G+C    CP   +  GTC   + kCC+
  gi|298772|     6    NFVTCRINRGFCVPIRCPGHRRQIGTCLGPRIKCCR    41

gi|5921174|sp|P46167|BD09_BOVIN: domain 1 of 1, from 20 to 55: score 6.4, E = 2.1
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                            C    G+C    CP   +  GTC     kCC+
  gi|5921174    20    NFVTCRINRGFCVPIRCPGHRRQIGTCLAPQIKCCR    55

gi|28628181|gb|AAO32801.1|: domain 1 of 1, from 27 to 62: score 6.0, E = 2.3
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                         + C q GG+C    C    K  GTC     kCC+
  gi|2862818    27    TSFSCSQNGGFCISPKCLPGSKQIGTCILPGSKCCR    62

gi|15826275|pdb|1E4Q|A: domain 1 of 1, from 3 to 35: score 3.5, E = 4.8
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                            C++sG  C    CP   K  GTC     kCCK
  gi|1582627     3    ---TCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCK    35

gi|9957108|gb|AAG09211.1|AF181951_1: domain 1 of 1, from 28 to 61: score 3.0, E = 5.5
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                       +  C +  G+C    CP  T i G C +    CCK
  gi|9957108    28    -KSDCFRKNGFCAFLKCPYLTLISGKCSRFH-LCCK    61
```

```
gi|9957108|gb|AAG09211.1|AF181951_1: domain 1 of 1, from 28 to 61: score 3.0, E = 5.5
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                       +  C +  G+C    CP  T i G C +    CCK
  gi|9957108    28    -KSDCFRKNGFCAFLKCPYLTLISGKCSRFH-LCCK    61

gi|9789929|ref|NP_062702.1|: domain 1 of 1, from 27 to 61: score 2.8, E = 5.8
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                            C+  G  C    CP   +  G C + K +CCK
  gi|9789929    27    NPITCMTNGAIC-WGPCPTAFRQIGNCGHFKVRCCK    61

gi|9971962|gb|AAG10514.1|AF288371_1: domain 1 of 1, from 27 to 61: score 2.8, E = 5.8
                    *->DhYkCvqsGGqCLySaCPiyTKiQGTCypgKakCCK<-*
                            C+  G  C    CP   +  G C + K +CCK
  gi|9971962    27    NPITCMTNGAIC-WGPCPTAFRQIGNCGHFKVRCCK    61
```

Figure 4.5: The beta-defensin profile was queried against the nr database and 12 sequences were extracted. Eleven sequences were beta defensin sequences.

Appendix 2

List of parameters of the Dragon Motif Builder program

Parameter	Explanation
Infile	input file
Outfile	output file
EMSearchOption	EM search option 1)EM1 2) EM2
RandomLimit	Random Peak scan coefficient: 10-100 recommented, higher value= long search time
motiflength	User specified motif length
EMmaxLength	Maximun length for motif, ONLY applicable for EM2
motifNum	number of motifs user wants
IterationThreshod	Maximun iteration for one search, program will terminate the search when exceeds the threshold
ICThreshold	Information content threshold, to maintain the result's IC quality. Vary 0 - 2
EMCriteria	EM eliminating criteria. 1-> Eliminate the identified motif patterns 2- >Eliminate the sequences which contain the sequences
revCompOption	0-> No reverval completement 1-> Reversal completement option
dirOption	0-> Forward strand search 1-> Inverse strand search
Selectpos	position segment analysis 0- full sequence length analysis 1-> Segment sequences anlysis
Startpos	Segment start position,
Endpos	Segment end position
EMThreshold	EM search threshold, vary 0-1
bgAnalysis	background analysis, 0-> no background analysis, 1-> analysis with internal generation background sequences, 2-> user induce background sequences, 3-> user specified the background sequences with the percentage 4-> user define the background sequence by their own data file
KeepZero	Remove the poor patterns from the group
nucleatideA	percentage of A NN in the background sequences 0-100
nucleatideC	percentage of C NN in the background sequences 0-100
nucleatideG	percentage of G NN in the background sequences 0-100
nucleatideT	percentage of T NN in the background sequences 0-100 NOTE: nucleatideA+nucleatideC+nucleatideG+nucleatideT = 100
appearOption	pattern appearance option 0-> Single 1-> Pair 2-> Single&Pair
pairDistance	pattern pair distance
MarkovModelorder	Markov Model order, recommented 3rd order
bgSeqFile	background sequence file
MarkovTable	Markov loop-up table
PlotGraph	graph plotting option 0-> No, 1-> Yes
EValue	background pattern appearance threshold
bgMaxlen	the background length that user specified
ContrastCoeff	contrast ratio btw the target and background, range from 0-1
PThreshold	p-value threshold range from 0 -1
controlOption	0-> no e and p value control, 1-> e value control, 2-> p value control, 3 -> both
EPIteration	number of iteration for the p & e value control before we relax the threshold condition
ERatio	number of relaxation coefficient for e value