ABSTRACT

LI, LI. Computational Modeling of Dose Response Relationship for Steroid Hormone Receptor-Mediated Gene Expression and Prediction of Androgen Response Element. (Under the direction of Dr. Melvin E. Andersen and Dr. Steffen Heber.)

Steroid hormone receptors are critical targets of both synthetic drugs used in hormone therapy and environmental endocrine active chemicals (EAC). Gene expression mediated by steroid hormone receptors was found to exhibit a non-monotonic dose response relationship. To further investigate this relationship, an ordinary differential equation-based computational model was formulated to examine the effect of EACs that display non-monotonic, rather than the typical monotonic, dose-response behaviors. Where the agonist ligand is an agonist, a U-shaped dose-response appears as a consequence of the inherently nonlinear process of receptor homodimerization. A higher degree of U-shaped dose-response curve modulation is effected by mixed-ligand heterodimers formed between endogenous and exogenous ligand-bound monomers. A novel mechanism for non-monotonic, particularly U-shaped, dose-response behaviors observed with specific steroid homologs is provided through this work. This mechanism will help in not only understanding how selective steroid receptor modulators work, but also in the improvement of risk assessment for EACs.

Another focal point in this research is on the statistical approaches to the identification of androgen response elements (ARE). The regulation of gene expression is largely influenced by the behavior of DNA-binding transcription factors. Several computational methods have demonstrated their ability to predict transcription factor binding sites (TFBS) in the gene promoter regions. Namely, Support Vector Machine (SVM), Hidden Markov Model (HMM), and Random Forest (RF) all summarize sequence patterns of
experimentally determined TFBS. In order to strengthen the prediction of putative AREs in the human genome, three statistical methods were explored, whose cross-validation results indicated that they all provided good sensitivity and specificity in identifying AREs, with an accuracy of at least 80%. It is the first time HMM, SVM, and RF have all been applied to the construction of ARE prediction models.

As a complement to the first two topics, an elucidation of androgen receptor-dependent gene regulatory networks was pursued. Understanding the underlying mechanism and dynamics of androgen receptor-regulated gene networks requires knowing the direct target genes whose expression levels are modulated by the androgen signaling pathway. Through the inspiration of the Arabidopsis transcription network, a systematic analysis of human genome gene upstream regions was undertaken with the goal of identifying potential androgen response regulatory elements through the occurrences of regular expression patterns. A number of interactions have been suggested between the AR and other target gene transcription factors as a result of the transcription network and sequence analysis of the functional targets.
Computational Modeling of Dose Response Relationships for Steroid Hormone Receptor-Mediated Gene Expression and Prediction of Androgen Response Element

by

Li Li

A dissertation submitted to the Graduate Faculty of
North Carolina State University
In partial fulfillment of the
Requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, NC
2007

Approved by:

Dr. Melvin E. Andersen

Dr. Steffen Heber
Chair of Advisory Committee

Dr. Barbara Sherry

Dr. Zhao-Bang Zeng

Dr. Hao Helen Zhang
Dedication

To my family
Biography

Li Li graduated from Peking Union Medical College in Beijing, China with a medicine degree in 1999 and then continued her education at Illinois Institute of Technology, where she received a M.S. in Computer Science in 2003. She was inspired to join the Bioinformatics Program at North Carolina State University in 2003 because of dual interests in human health and computational biology. While working towards her doctoral degree, she interned at the Hamner Institutes for Health Sciences under the direction of Dr. Melvin E. Andersen.
Acknowledgments

The road through my graduate studies at North Carolina State University has been a mixture of long straight stretches, sharp turns, several rough bumps, hills, and finally a grand overlook. Navigating this road would have been impossible without the help of a number of people. I would like to thank, first and foremost, my supervisor at the Hamner Institutes for Health Sciences, Dr. Melvin E. Andersen, and my advisor, Dr. Steffen Heber for their continual patience and guidance during my doctoral studies. I am also thankful for their support and advice in my research. They have always been patient and provided encouragement and constructive thoughts that helped me bring this dream to fruition. Next, I am grateful for the assistance and direction from Dr. Qiang Zhang, who answered many questions and clarified several points of ambiguity in my work. I would also like to thank my committee members for embracing the multidisciplinary nature of my research: Dr. Zhao-Bang Zeng, for his great direction and support over my studies, Dr. Barbara Sherry for her kindness and valuable insight, and Dr. Hao Helen Zhang for her statistical expertise.

In addition to my advisory committee, I would like to thank everyone at Hamner Institute for their reassurance and support during these long years. Further, I would like to thank Juliebeth Briseno for her assistance with many technical administrative issues, as well as fellow graduate students Chun-Ju Wang, Chuanhua Xing, and John Cromer for their immense emotional support during my time at NCSU. The challenges and struggles
I faced were made much less painful by these kind souls who were always ready to lend an open ear, a kind heart, and a helping hand.

Of course, none of this would be possible without the support of my family, including my parents and my brother. They always believed in me, my abilities, and my dreams. I am very fortunate and grateful to have such a supportive family during both bright and dark hours of my life.
# Table of Contents

List of Tables .................................................................................................................x

List of Figures.................................................................................................................xi

## 1 Review ......................................................................................................................1

1.1 Steroid Hormone Receptor ..................................................................................2

1.1.1 The Basics of Steroid Hormone Receptors ...............................................2

1.1.2 Steroid Hormone Receptor Coregulatory Proteins .....................................4

1.2 Endocrine Active Chemicals ..............................................................................7

1.3 Mathematical Modeling of Dose Response Curves .......................................8

1.4 Androgen Receptor and Androgen Response Element .............................10

1.5 Machine Learning Approaches for Prediction ...........................................13

1.5.1 Hidden Markov Model .............................................................................14

1.5.2 Support Vector Machine ..........................................................................16

1.5.3 Random Forest .........................................................................................17

1.6 Reference ...........................................................................................................18

## 2 Statistical Approaches to Identify Androgen Response Elements ....23

2.1 Abstract ............................................................................................................24

2.2 Introduction .....................................................................................................25

2.3 Methods ..........................................................................................................27

2.3.1 Data Collection ........................................................................................27

2.3.2 Feature Selection .......................................................................................27

2.3.3 Construction of Hidden Markov Model .................................................29
2.3.4 Support Vector Machine .............................................................. 30
2.3.5 Random Forest ........................................................................... 31
2.3.6 Performance Assessment ............................................................... 31
2.3.7 Genome-wide Identification of Putative AREs ................................. 32

2.4 Results .......................................................................................... 32
2.4.1 HMM Model Architecture ............................................................. 32
2.4.2 HMM Model Evaluation ................................................................. 33
2.4.3 SVM Model Evaluation ................................................................. 34
2.4.4 Comparison of HMM, SVM and RF models ...................................... 35
2.4.5 Genome-wide Identification of AREs .............................................. 36

2.5 Discussion ................................................................................... 36

2.6 Acknowledgements ....................................................................... 40

2.7 Reference ..................................................................................... 41

3 Androgen Receptor-Dependent Gene Regulatory Networks ............ 56

3.1 Abstract ....................................................................................... 57
3.2 Introduction .................................................................................. 58

3.3 Methods ....................................................................................... 60
3.3.1 Microarray Data Analysis ......................................................... 60
3.3.2 k-mer distributions ................................................................. 61
3.3.3 Regulatory Sequence Analysis ................................................ 62
3.3.4 F-score Selection ...................................................................... 63
3.3.5 Random Forest Prediction ....................................................... 64
3.3.6 Functional Annotation.....................................................................................65
3.4 Results..................................................................................................................65
  3.4.1 Differentially Expressed Genes......................................................................65
  3.4.2 k-mer Feature Selection..................................................................................66
  3.4.3 Regulatory sequence analysis.........................................................................66
  3.4.4 Prediction Results on Training Data Set.........................................................67
3.5 Discussion.............................................................................................................69
3.6 Acknowledgements..............................................................................................72
3.7 References...........................................................................................................72

4 Non-monotonic Dose Response Relationship in Steroid Hormone Receptor-Mediated Gene Expression.........................................................85
  4.1 Abstract..............................................................................................................86
  4.2 Introduction.........................................................................................................87
  4.3 Methods.............................................................................................................89
    4.3.1 Model Structure............................................................................................89
    4.3.2 Model Parameters.......................................................................................92
    4.3.3 Modeling Tools............................................................................................92
  4.4 Results................................................................................................................93
    4.4.1 Exogenous Ligand X as a Pure Agonist.......................................................93
    4.4.2 Exogenous Ligand X as an Antagonist.......................................................98
    4.4.3 Exogenous Ligand X as a Partial Agonist...............................................100
  4.5 Discussion...........................................................................................................101
  4.6 Acknowledgements............................................................................................107
List of Tables

Table 2.1 Summary of experimentally verified known androgen response element ..............45

Table 2.2 F-score selection of 4-mer motifs in AR-regulated genes ................................52

Table 2.3 Summary of the performance ........................................................................53

Table 2.S Supplemental Table ......................................................................................54

Table 3.1 F-score selection of 4-mer motifs in AR-regulated genes ............................78

Table 3.2. Top ranked TFBS names, recognition sequences, and their corresponding F-score values in order of importance .................................................................81

Table 3.3. The prediction results using different k-mers .............................................82

Table 3.4. Performance comparison of the RF classifier of TFBS with of combination of TFBS and mismatched 5-mer features in DHT-regulated gene expression .........................83

Table 3.5. Comparison of mismatched 5-mer features and the TFBS motifs ............84

Table 4.1 Conditions in which U- or inverted U-shaped dose-response curves may appear ........................................121

Table 4.S1. Species, initial values, and ODEs for the SHR-mediated gene expression model 122

Table 4.S2. Default reaction constant values for the SHR-mediated gene expression model 125

Table 5.1. Top ranked TFBS names and recognition sequences, along with their corresponding motifs ........................................................................................................136
List of Figures

Figure 1.1 Functional domains of steroid hormone receptor ............................................... 3

Figure 2.1 Graphical representation of HMM framework for AR binding sites prediction........ 46

Figure 2.2 General SVM prediction scheme ........................................................................... 47

Figure 2.3 A graphic representation of an aligned set of AR binding site ......................... 48

Figure 2.4 Performance of HMM model ................................................................................. 48

Figure 2.5 Performance of SVM model ................................................................................. 49

Figure 2.6 A contour plot of the fitted decision values for ARE classification ................. 50

Figure 2.7 Performance of SRF model .................................................................................. 51

Figure 3.1 Random forests classification process ................................................................. 75

Figure 3.2 The scatter plot results after display options have been set and up-regulated and down-regulated genes selected .......................................................... 76

Figure 3.3 F-scores are used for feature selection for mismatched 5-mer dataset ............ 77

Figure 3.4 Screenshot of Toucan 2 with known TFBS analysis result ................................ 79

Figure 3.5 Curves of F-scores against TFBS features .......................................................... 80

Figure 4.1 Structure of the model for SHR-mediated gene expression ............................. 112

Figure 4.2 U-shaped steady-state dose response in gene expression and magnitude of U-shape when exogenous ligand X is a pure agonist and the heterodimer LRRX is absent ............... 113

Figure 4.3 Linearization of the homodimerization process eliminates the U-shaped dose response when exogenous ligand X is a pure agonist and the heterodimer LRRX is absent ............... 114

Figure 4.4 Mixed-ligand heterodimer LRRX acting as an activator reduces and inverts the U-shaped dose response when exogenous ligand X is a pure agonist ............................... 115

Figure 4.5 Mixed-ligand heterodimer LRRX acting as a repressor deepens the U-shaped dose response when exogenous ligand X is a pure agonist ............................... 116
Figure 4.6 Monotonically decreasing dose response in gene expression when exogenous ligand X is a pure antagonist and the heterodimer LRRX is absent

Figure 4.7 Mixed-ligand heterodimer LRRX acting as an activator rightward shifts and converts to inverted U-shape, the monotonically decreasing dose response when exogenous ligand X is a pure antagonist

Figure 4.8 Mixed-ligand heterodimer LRRX acting as a repressor leftward shifts the monotonically decreasing dose response when exogenous ligand X is a pure antagonist

Figure 4.9 Effect of CoR/CoA ratio on the dose response when exogenous ligand X is a partial agonist

Figure 5.1 Schematic representation for the mechanisms of action of androgen

Figure 5.2 Venn diagrams summarize the number of differentially expressed genes at different time points

Figure 5.3 Distribution of transcription factor binding sites common in ARE positive gene groups

Figure 5.4 Distribution of transcription factor binding sites common in ARE negative gene groups
Chapter 1

Review
1.1 Steroid Hormone Receptor

In the early 20th century, small lipophilic hormones were isolated based on their abilities to affect development, differentiation, metamorphosis, and physiology in humans (Mangelsdorf 1995). They diffuse directly across cell plasma membranes and bind to intracellular receptors. These receptors are structurally related, constituting the so-called steroid hormone receptor (SHR) superfamily, and are activated after binding of their specific ligands (Kawata 2001). SHRs have been traditionally considered to act via the regulation of transcriptional processes, involving nuclear translocation and binding to specific response elements, and ultimately leading to regulation of gene expression. It has now been 20 years since the isolation of cDNAs encoding glucocorticoid and estrogen receptors, which were the first cellular transcription factors to be cloned (Hollenberg 1985; Green 1986). Following the work came the identification of receptors for all known SHRs such as androgen receptor and the emerging concept that identifies target genes regulated by these receptors. Identification of SHR family members, including estrogen receptor (ER), androgen receptor (AR), and progesterone receptor (PR), as well as the glucocorticoid receptor (GR), suggested the likely universal nature of this superfamily.

1.1.1 The Basics of Steroid Hormone Receptors

SHRs share a common structure of four major domains that are highly conserved: DNA binding domain (DBD), ligand binding domain (LBD), a N-terminal transactivation domain (NTD), and hinge region (Figure 1.1). Most of the steroid hormone-dependent phosphorylation sites on SHRs are located in the NTD, which is critical for SHR function because it contains the major transactivation region referred to as Activation Function 1 (AF1). NTD exhibits little homology among SHRs and is responsible for ligand-induced transactivation together with LBD regions in SHRs (Rochette-Egly 2003). The well-conserved DBD can fold into two zinc-coordinated finger structures with the ability to bind to DNA. All members of SHRs mediate the expression of a gene by binding to hormone responsive elements (HREs) as dimers, with each monomer recognizing
the individual DBDs. The carboxyl-terminal LBD is organized into 12 α-helical folds that create a hydrophobic cavity for hormone binding. Upon ligand binding, helix 12 reorients to a position on the LBD and is responsible for the formation of a hydrophobic groove, which allows for the docking of coregulators to the LBD surface, leading to transcriptional activation by receptors (Mahajan 2005). LBD also contains a ligand-dependent nuclear translocation signal, determinants to bind chaperone proteins, dimerization interfaces, and a potent ligand-dependent activation domain referred to as activation function 2 (AF-2) (Klein-Hitpass 1998). The hinge region facilitates the subcellular transport of SHRs from the cytoplasm to the nucleus.

Prior to ligand binding, the inactivated SHRs are thought to be localized in the cytoplasm and associated with a chaperone complex composed of immunophilins, heat shock protein 90 (HSP90), and other factors (Pratt 1997). The chaperone complex can facilitate the folding of the receptor to enable it to bind ligands, stabilize it in a conformation receptive to ligand binding, and alter the affinity of steroid receptors for their ligand. Upon ligand binding, SHRs are activated and transported to the nucleus from the cytoplasm. Binding of the cognate hormone induces a cascade of events including heat shock protein dissociation, conformational changes, phosphorylation, and dimerization, which ultimately results in forming SHR dimers that are able to bind to well-defined HREs. The distinct conformation of the SHR that is induced by ligand binding reflects the size and shape of the ligand, and determines the agonist or antagonist nature of the response (Katzenellenbogen 2003). After SHRs are occupied with agonists/antagonists, the liganded SHRs binding to target genes requires ligand-dependent SHR dimerization. Protein-protein interaction studies show that the interaction between NH₂- and carboxyl-terminal domains,
which contain the LBD, raise the possibility of forming SHR dimers (Langley 1995). This is because most nuclear receptors have only a limited intrinsic DNA-binding affinity, but the homodimer is around 100-fold more efficient than the binding of just one of its subunits to an isolated half-site. Therefore, SHR usually binds to its specific HRE as a homodimer, with each monomer recognizing one half-site of the response element. For example, core estrogen response element sequences are arranged as an inverted repeat (AGGTCANNTGACCT) or direct repeat (AGGTCANNNAGGTCA) with 1-5 spacer nucleotides separating the half-site (Whitfield 1999; Klinge 2001), and are highly conserved across species. After binding to their specific HREs, the receptor-dimer complexes then recruit the basal transcriptional factors to form the complex that transcribes a gene, or more likely groups of genes, throughout the genome.

1.1.2 Steroid Hormone Receptor Coregulatory Proteins

The fact that SHRs regulate transcription through interaction with HREs that function as enhancer elements at a variable distance from the promoter, suggests the existence of other mechanisms for communication between SHRs and basal transcriptional machinery. Evidence for the existence of coregulator proteins originally came from studies involving transcriptional squelching or interference (Tora 1989). The receptor dimer-DNA complex can interact with a battery of nuclear protein coregulators that enhance or reduce transactivation of target genes, but do not significantly alter the basal transcription rate. Coregulators, which do not possess specific DNA binding affinity, can be divided into two major types: (1) coactivators directly regulate the receptor transactivation function, which may act as a bridge between the DNA-bound receptors and the transcriptional machinery, or those that harbor specific enzyme activities, such as histone acetyltransferases or methyltransferases; (2) corepressors, which may block the transactivation domain or recruit complexes with histone deacetylase activity, mediate receptor-dependent repression of transcription (Lavery 2005). In general, coregulator proteins interact with DNA indirectly through SHRs, and these coregulators participate in DNA modification of target genes,
either directly through modification of histones or indirectly by the recruitment of chromatin-modifying complexes (Heinlein 2002). Many coregulators are most likely recruited at the promoter of these receptor-DNA complexes (McKenna 2002). When present on target promoters, transcriptional coregulators play different roles either due to their specific enzymatic activities or their ability to recruit other regulator proteins. Coactivators, SRC-1 and CBP/p300 possess histone acetyltransferase activity and may modulate SHR activity via their ability to remodel chromatin structure, and maintain a transcriptionally open chromatin structure at the promoter of HRE (Rahman 2004). Conversely, SHR corepressors play a crucial role in transcriptional repression, and their function in part requires recruitment of histone deacetylases. For instance, SHR corepressors SMRT and NcoR function as part of larger protein complexes that include histone deacetylases, which enhances nucleosome-DNA interactions and inhibits transcription factor recruitment and gene expression (Smith 2004).

Although the molecular basis of the interactions between SHRs and coregulators was not well defined, the identification of some SHR coactivators and corepressors has opened a pathway to a greater understanding of hormone action. Hypothesized interaction between coactivators and SHRs in an agonist-dependent manner was promoted by agonist and inhibited by antagonist in the initial study of coactivator SRC-1, ER, and PR (Onate 1995). On the contrary, corepressors bind to SHRs in the presence of their respective antagonists. Transient transfection assays (Jackson 2000) demonstrated that NCoR and SMRT selectively repressed the agonist activity and induced active repression by promoting their association with negative regulation of SHR activity. It has also been accepted that antagonists exert their effects by a passive mechanism driven mainly by the ability of antagonists to compete with agonists for the ligand binding site. However, recent analyses of antagonist-occupied receptor functions (Schulman 1996) suggest antagonists are involved in active recruitment of corepressor or coactivator proteins to produce mixed transcriptional activity. The competitive equilibrium binding of agonist and/or antagonist
may control the direction and the rate of gene transcription. Surprisingly, some antagonists inhibit SHR activity in a selective manner and may even activate transcription under certain conditions (Katzenellenbogen 1996). For example, SMRT abrogates the ability of mixed antiestrogen to activate transcription of an ER-dependent gene (Smith 1997). Receptor ligands that exhibit agonistic or antagonistic biocharacter in a cell- and tissue context-dependent manner are defined as selective receptor modulators (SRMs). These ligands act either as partial agonists or antagonists, in a manner contingent upon the tissue or promoter contest (Mckenna 1999). In women, tamoxifen shows tissue-selective difference, with strong agonist activity indicated by uterine cell proliferation, but with little stimulation of the breast (Evans 1993). The possible mechanism underlying tissue selectivity could be tissue-specific recruitment of coactivators and corepressors during the SHR signaling pathway. The existence of a cellular equilibrium of coactivators and corepressors can be shifted toward corepressor preference by antagonist. Therefore, SRM is determined at least in part by the specific set and/or the ratio of coactivators and corepressors expressed in a given cell or tissue. AF-1 and AF-2 are the major regions to recruit both coactivators and corepressors. The interaction of the SHR-AF1 domain with general transcription factors induces folding AF-1 domain interacting with coactivators and provides an interaction surface for the SHR COOH terminus, and this in turn enhances subsequent protein-protein interactions (McEwan 2004). In the mean time, hormone bindings induce a conformational change in AF-2 that creates a new protein interaction site on the surface of the LBD that can be recognized by coactivators such as p160 family. However, chromatin immunoprecipitation (ChIP) assays suggest that receptor and coregulator association in gene promoters is temporally regulated, and the SHR transcription complex appears to repeatedly bind onto and off of target promoters in the presence of continuous stimulation by steroid hormone (Shang 2000). Observation of a cyclic association and dissociation of coactivators with the promoter, occurring in opposite phases, suggests an exchange between these coregulator complexes at the target promoter (Burakov 2002). Accordingly, the occupancy of the LBD, and therefore its conformation
change determines whether SHRs interact with coactivators or corepressors, and then activate or repress transcription. The ability of SHR to activate transcription is a product of the ability of the receptor to interact with coregulators and other proteins required for gene expression, and the effect of various enzymatic activities on the formation, function, and disassembly of the receptor-coregulator complex. The actions of SHRs through their regulation on target gene expression (termed genomic actions) are critical for the control of many biological processes. Non-genomic action of SHRs refers to their role in regulating rapid cellular signaling transduction that does not require changes in gene expression regulated by the SHR-hormone complex binding to DNA-response elements for the SHR (Aranda 2001). For instance, estrogens activate Mitogen-Activated Protein Kinases (MAPK) or Tyrosine Kinases-dependent pathways which can stimulate growth of breast cancer cells, independent from gene transcription. Understanding the molecular mechanism is therefore important, and may in the future turn out to be of relevance for clinical purposes.

1.2 Endocrine Active Chemicals

SHRs that normally respond to endogenous ligand, however, can also respond to exogenous substance, both of natural or man-made origin. Chemicals in the environment that interact directly with steroid receptors as agonists or antagonists have the potential to produce specific reproductive and developmental effects at extremely low concentrations. The Environmental Protection Agency (EPA) defined an endocrine active chemical (EAC) (also called endocrine disruptor ED), as “an exogenous agent that interferes with the synthesis, secretion, transport, binding action, or elimination of natural hormones in the body that are responsible for the maintenance of homeostasis, reproduction, development, and/or behavior”. Substantial evidence has surfaced on the hormone-like effects of environmental chemicals such as pesticides and industrial chemicals. Since these exogenous substances possess significant hormonal activity, the endocrine and reproductive effects of these EACs are believed to be due to their ability to
mimic the effects of endogenous hormones, antagonize the effects of endogenous hormones, and disrupt the synthesis and metabolism of endogenous hormones (Mendes 2002). One piece of the available evidence is that DDE, a metabolite of a major pesticide, has been found to bind to AR and block testosterone-induced cellular response in vitro. Studies also reported that high levels of exogenous estrogen DDE or PCBs might be a contributing factor for breast cancer in women (Lemaire 2006; Prince 2006). In addition to these studies showing that synthetic compounds can interact with the SHRs, plant oestrogens affect animal reproduction and sexual differentiation in vivo (Levy 1995). These findings may have broad implications concerning how EACs could interact with SHR pathway but bring about effects different from those caused by endogenous hormones.

It is important to understand the potential human health implications of exposure to environmental chemicals that may act as EACs. It is also necessary to have an understanding of how pharmaceutical and personal care products and other chemicals affect human diseases and how they may correlate. To provide more accurate guidance for health risk assessment of EACs, and recommend appropriate preventive and protective strategies, It is important to understand the underlying cellular and molecular mechanisms of these complex dose-response relationships. Dose response assessment is an integral step in evaluating potential health risks to human and animals at exposure levels of interest.

1.3 Mathematical Modeling of Dose Response Curves

The deleterious effects of hormone mimics on the endocrine system can be detected after the chemicals are released in the environment. These EACs usually cause dose-dependent toxic effects for most of the response they induce. However, it has been widely accepted that the most fundamental, common and generally applicable dose response model is the hormetic U-shaped or inverted U-shaped biphasic model rather than the threshold and linear models which had been
dominant for decades (Calabrese 2001). Hormesis, either a U shaped or inverted U shaped dose response, becomes a critical determinant affecting the nature of dose response relationships, especially in the area of low or modest disruption. Hormetic effects were observed at low exposure levels based on the dose response pattern with data from developmental studies, indicating that there might actually be a reduced risk of toxic effects at low exposure levels (Hunt and Bowman 2004). Understanding effects of low dose is a necessary part of understanding the cause and effect of relationship between chemical exposure and illness. Assessment of the published literature indicates there is no single mechanism that accounts for the hormetic-like biphasic dose responses. The dose responses are mediated via different agonists and receptors depending on the tissue, cell type and endpoint, indicating that the phenomenon of hormesis might result from multiple mechanisms and therefore comprise diverse biological processes. However, a general explanation for the hormesis phenomenon is that at low levels of disruption or toxicity, many biological systems display an overcompensation response, which results in the apparent low-dose stimulation component of the response curve, while at higher doses with greater initial toxicity, the system often displays a more limited capacity for a compensatory response, usually insufficient to return to control values (Calabrese 2004). The overcompensation originates from activated defense mechanisms and leads to possibly beneficial low dose effects. Therefore, therapeutic activities and environmental regulatory actions can be based on low dose extrapolation. They are expected to have potential beneficial effects in preventing steroid hormone imbalance diseases such as breast cancer, osteoporosis, and cardiovascular disease. For example, the use of estradiol in low doses is an effective breast cancer treatment strategy, but not at high doses. Improvement of risk assessment strategies may include taking into account biphasic dose response curves and hormesis.

Computational modeling is an indispensable tool for understanding a dose response relationship of SHR-induced gene expression by agonist or antagonist, as the structural complexity of the
networks and the richness of their possible behaviors can only be understood with more rigorous mathematical analysis. The study in chapter 4 is intended to dissect the effects of individual components of the network that controls the SHR-mediated gene expression. We develop a specific mathematical model that captures the most important aspects of the processes and a numerical implementation of this mathematical model to qualitatively and quantitatively predict dose-response behaviors of SHR-mediated gene expression. Differential equation models provide a general framework in which gene regulation processes are considered. We then transform essentially a system of chemical kinetic reactions into a system of nonlinear first-order ordinary differential equations (ODEs), representing rate changes in concentrations of mRNA, proteins and so on. An ODE describes the evolution over time of the concentration of reactants and products. To integrate the system of differential equations, one needs not only the equations themselves but also a set of initial conditions—the value of concentrations at time zero. Then the concentrations of each chemical species change over time according to differential equations where the time evolution of the system of reactions may be regarded as continuous and deterministic. Equilibrium states, if any, in the differential equations will be reached when there is no further change in state. Only steady state values are used to construct the dose response curves. Therefore, these biochemical events as well as ligand binding, receptor dimerization, and DNA response element occupation and so forth, can be translated into a mathematical model for quantitative risk assessment of human health effects.

1.4 Androgen Receptors and Androgen Response Elements

Androgen receptor (AR), a member of the SHR family functioning as a ligand-activated transcription factor, mediates male sexual differentiation, pubertal sexual maturation, and other male reproductive functions. In addition to regulating normal physiological and pathological development, AR may also have an important impact on abnormalities associated with the developing male reproductive system (Kelce 1997). For instance, in aging men, the AR not only
mediates prostate development but also serves as a key regulator of primary prostatic cancer growth. Unbound testosterone in circulation diffuses and enters the prostate. Within the prostate, dihydrotestosterone (DHT), one of the principal natural androgen, binds to the AR with a high affinity to form an intracellular DHT-AR complex, which then binds to the androgen response elements (ARE) in target genes, ultimately inducing DNA synthesis and cellular proliferation (Hsing 2002). AR has also been implicated in several other cancers that have a hormonal basis, including liver, breast, ovarian and endometrial cancers. Gene expression is often used to measure AR activity in cells and cancer progression in patients. Close examination of the androgen-regulated gene expression may provide clues as to how the entire gene regulatory network is mediated by androgen. Understanding the pathways and genes regulated by AR is critical to the design and implementation of effective therapies for androgen-dependent diseases.

In general, ligand-activated AR can modulate gene expression directly by interacting with specific elements in the regulatory regions of various target genes through DBD (Nazareth 1996). Upon androgen binding, AR is phosphorylated and forms a homodimer, which is transported to the nucleus, where it mediates transcription through interacting with specific AREs present in the promoters of target genes (Keller 1996). The sequence of ARE is highly conserved across species, and the region is composed of two palindromic hexanucleotide half-sites that are separated by a three-nucleotide spacer (Roche 1992). AR binds as homo- or heterodimers to these hexamers, each dimer partner interacting with one of the hexamers (Luke 1994). In addition, AR-responsive genes often contain enhancer and promoter elements in their regulatory regions. Biochemical and genetic studies revealed that both the enhancer and the promoter displayed androgen responsiveness, but maximal activity required the presence of both (Shang 2002). Small differences in the half-site sequences and the sequences of the flanking extension of the response elements also appeared to be important parameters in determining receptor binding efficiency, resulting in different levels of activation of AR-mediated transcription.
The classical view of AR binding to recognition sites is that hormone-receptor complex remains at the binding site for as long as the ligand is present in the cellular milieu. Alternatively, the receptor may interact transiently with a response element, recruiting a secondary set of factors that in turn form a stable complex at the regulatory site (McNally, Muller et al. 2000). A rapid and transient interaction of AR with androgen response elements both in vivo and in vitro, supports the dynamic "hit-and-run" model in which AR first binds to chromatin after ligand activation, recruits a remodeling activity, and is simultaneously lost from the template. (Schaffner 1988). Androgens regulate gene expression through several mechanisms (Zhou 2005). First, primary response genes are transcriptionally activated or suppressed rapidly after binding of ligand-activated AR to androgen response elements in their 5'-flanking regulatory regions. Second, the delayed primary response genes with AREs respond to androgen stimulation via AR binding, but the maximal response requires interaction with cofactors and synthesis of other ongoing proteins. Third, secondary response genes without AREs also can respond indirectly to androgen stimulation via activation or suppression of transcription by other factors directly regulated by AR. This response requires protein synthesis and, therefore, has a temporal delay.

Identification and characterization of consensus AREs to evaluate their hormonal transcription is critical to understanding the possible mechanism of androgen-specific gene regulation. We use genome-wide approaches combined with computational analysis in identifying hormone-responsive genes. The objective of Chapter 2 is to characterize the temporal program of transcription that reflects cellular responses to androgens and to identify specific androgen-regulated genes or gene networks that participate in these responses. It may contribute to the systematic elucidation of the gene regulatory network mediated by androgen, which will be pivotal for the development of treatment for androgen-related diseases, such as prostate cancer. Endocrine therapies may offer the potential for prevention of the activation of AR that leads to transcription of genes involved in regulation of cell cycle and survival.
1.5 Machine Learning Approaches for Prediction

The transcription of genes is controlled by transcription factors (TF) which bind to short DNA sequences known as transcription factor binding sites (TFBS). With the availability of genome sequences for multiple species and large-scale gene expression data, computational predictions of TFBSs in eukaryote genomes have been made significant advances in the past 10 years. The prediction of TFBSs is of importance for unraveling the underlying molecular mechanisms of biological processes. Many methods for identifying target sites for a TF have been proposed. All learning techniques fall into two broad classes: supervised and unsupervised. The goal of supervised learning is to design a system able to accurately predict the class membership of new objects based on the available features. In contrast, the aim of unsupervised learning is to explore the data and discover the similarities between objects (Tarca 2007). The main difference between them is that the first effectively makes use of prior knowledge in order to achieve accurate results, whereas the latter is constructed in such a way that it does not need additional, previously derived information about the data to be used (Soinov 2003). Supervised learning methods construct a decision rule from a training set of known positive and negative examples and algorithms such as Support Vector Machine learning to discriminate between training examples from each category. Unsupervised methods begin with less well-defined information, for example a set of co-expressed genes from a microarray study which are thought to contain some set of common but unknown patterns. A wide range of machine learning methods have been considered to better organize data, to infer regulatory relationships of genes, to discover disease subtype signatures, and to derive various predictions.

There are numerous successful applications of machine learning in biology abound. For instance, gene expression data was successfully applied to classify patients in different clinical groups and to identify new disease groups (Perou 1999; Alizadeh 2000), while genetic code allowed
prediction of the protein secondary structure (Tarca 2005). Unique physical properties of machine learning algorithms were applied to microarray data to predict TF-target association (Holloway). Identifying TFBSs is helpful for understanding the mechanism of transcriptional regulation. Since TFs, typically 10-15 nt in length, are short and occur very frequently in larger genomes, it is not easy to recognize them. However, binding sites are often preserved through evolution, and thus become apparent. The abundance and the diversity of genomic data provide an excellent opportunity for identifying TFBSs. Developing methods to integrate various types of data has become a major trend in this pursuit (Tsai 2006). A general approach is to formulate a mathematical representation of the binding pattern of a given factor based on collections of confirmed binding site sequences. The representation is subsequently used to score candidate sequences for occurrences of said pattern (Georgi 2006). Positional weight matrix (PWM) is a commonly used representation of identifying TFBS. The binding profile for a TF matrix is calculated by aligning identified sites and then counting the frequency of each DNA base at each position of the alignment. According to PWM model, each base of the site contributes independently to the binding of the transcription factor (Vavouri 2005). Several hundred matrices for specific transcription factors are available through the databases TRANSFAC (Wingender 1996) and JASPAR (Sandelin 2004). Here we review the following three popular statistical approaches, Hidden Markov Model, Support Vector Machine, and Random Forest to identifying TFBSs.

1.5.1 Hidden Markov Model

A Hidden Markov Model (HMM) is a statistical model where hidden parameters are determined from observable parameters in a system governed by a Markov process with unknown parameters. HMMs form the foundation of many bioinformatics tools employed in feature or sequence function prediction. Information from the past, known as training data, can be incorporated into a HMM model to derive accurate predictions of future responses. There are
several common tasks that can be performed with the assistance of a HMM model. One task is computing the probability of a particular output sequence, given the parameters of the model. This can be accomplished using Forward-Backward algorithm. Another task is finding the most likely sequence of hidden states that could have generated a given output sequence, given the parameters of the model. The Viterbi algorithm is utilized to accomplish this task. A third common task is to find the most likely set of state transition and output probabilities, given an output sequence or a set of output sequences. The Baum-Welch algorithm is often used for this purpose. In general, a HMM model applies Bayesian statistics to estimate the true frequency of a residue at a given position in the alignment from its observed frequency. HMM is typically defined as a stochastic finite state automaton that is built up from a finite set of possible states. An HMM model consists of a set of states, where each state can emit symbols (nucleotides or amino acids) based on a probability distribution. The emission probability for a certain nucleotide is specific for each state. States are connected in a chain-like structure, where the probability of moving from one state to another is termed a transition probability (Sandelin and Wasserman 2005). It is convenient to think of a HMM as a model that generates sequences. Starting in an initial state, we choose a new state with some transition probability, either staying in state 1 with transition probability $t_{1,1}$, or moving to state 2 with transition probability $t_{1,2}$; Then we generate a residue with an emission probability specific to that state. Briefly, the fundamental equation relevant for this process is a restatement of Bayes’ rule. Let $x_i$ represents state at $i = 1, \ldots, N$ and $y = y_1, y_2, \ldots, y_N$ that maximizes $Pr(x, y)$. Notice that the path that maximizes $Pr(x, y)$ must be identical to the path that maximizes $Pr(y|x) = \frac{Pr(x, y)}{Pr(x)}$. The path through the Hidden Markov Model maximizes $Pr(x_{i+1}, y_{i+1}) = \max \Pr(x_{i+1}, y_{i+1} \mid x_i, y_j) \Pr(x_i, y_j)$, where $x_i$ includes states $x_1 \ldots x_i$, and $y_i$ is the path $y_1 \ldots y_j$ that maximizes $Pr(x_i, y_j)$. HMM models are constructed with a description of the variation in the features of interest. The structure and parameters of the
model are obtained through a training phase. The models then apply the above algorithm to the new sequence to find the most probable path of hidden states that emitted the observed sequences. The most probable path is then compared against the known promoter features to detect potential binding sites. Predictions are calculated by multiplying the probabilities of the observed symbols being emitted times the probabilities of transitioning from the respective states to respective states. Identifications of sequence motifs in a genome are often accomplished with the aid of a profile HMM Model.

1.5.2 Support Vector Machine

Support Vector Machine (SVM) is one of the learning machines that can perform classification by constructing a large-dimensional hyperplane that optimally separates the data into two categories. SVM has been widely applied within the field of computational biology to pattern recognition problems, such as microarray gene expression analysis, prediction of protein-protein interactions, and sequences analysis. It comprises a number of similar methods of supervised learning used for both regression and classification. One hallmark feature of SVMs is that they maximize the geometric margin while simultaneously minimizing the empirical classification error (Scholkopf 2004). For this reason, SVM is commonly referred to as a maximum margin classifier. SVM classification algorithms, proposed by Vapnik (Vapnik 1998) to solve two-class problems, are based on finding hyperplanes separating classes of training data. Every data point may be represented by a $p$-dimensional vector, and can be put in only one of two classes. The goal is to separate these points with a hyperplane containing $p-1$ dimensions, a structure typical of a linear classifier. However there may be a large number of linear classifiers that satisfy this requirement. The optimum solution is to maximize the margin between the two classes. In other words, it is important to choose the hyperplane such that the distance between the hyperplane is maximized. The nearest data points along the hyperplane are called support vectors. The algorithm works in such a way: given training data consisting of $N$ pairs $((x_i, y_i))$, $i = 1, ..., N$, the SVMs require the
solution of the following optimization problem by finding a maximized hyper plane \((W, b)\):

\[
\min_{w,b,\xi} \frac{1}{2}w^Tw + C\sum_{i=1}^{N} \xi_i \quad \text{subject to } \ y_i(w^T\phi(x_i)+b) \geq 1-\xi_i, \xi_i \geq 0.
\]

Input vectors in SVMs are mapped to higher dimensional spaces by function \(\phi\), where a maximal separating hyperplane is achieved. It is assumed that the maximal distance solved by the above equation will result in the best classifier generalization error. Furthermore, \(K(x_i, x_j) = \phi(x_i)^T\phi(x_j)\) is a kernel function. Though new kernels are being proposed by researchers, the following four basic kernels are used in SVM classification, where \(\gamma\), \(r\), and \(d\) are kernel parameters.

- **linear**: \(K(X_i, X_j) = X_i^T X_j\)
- **polynomial**: \(K(X_i, X_j) = (\gamma X_i^T X_j + r)^d, \gamma > 0\)
- **radial basis function (RBF)**: \(K(X_i, X_j) = \exp(-\gamma\|X_i - X_j\|^2), \gamma > 0\)
- **sigmoid**: \(K(X_i, X_j) = \tanh(\gamma X_i^T X_j + r)\)

In contrast to most machine learning methods, SVMs can handle non-vector inputs easily, as in the case of variable length sequences or graphs. Data types such as these are common in biological applications, and often are dependent on the construction of knowledge-based kernel functions (Noble 2003). Unlike correlation-based approaches, which consider the significance of individual features, the SVM considers the significance of a feature in the context of the features already selected, which may be useful in considering the effects of combinations of features on gene expression.

### 1.5.3 Random Forest

Random forests (RF) are machine learning based classifiers comprising decision trees. The method is originally derived from random decision forests, first formulated by Tin Kam Ho. Breiman combined the “bagging” concept and the “random subspace method” together to yield a collection of decision trees with controlled variations (Breiman 2001). The general idea of
applying RF approach is to construct a series of single classification trees. Each individual tree is built from the original data set by first sampling the number of cases at random, with replacement. Next, a number of variables much smaller than the number of input variables are selected such that the best separation of samples is used to split the node. This number is used repeatedly to increase the size of the forest by adding more trees. With no pruning, every tree is grown as large as possible. Finally, new objects can be classified by placing input vectors down each of the trees in the forest. The classification decision could be determined by the majority of votes over all trees in the forest.

To assess the performance of prediction models, estimating the generalization error is a very important step in machine learning techniques. The error rate for the RF model is dependent on two key factors: the correlation between two trees in the forest, and the strength of each individual tree in the forest. An increase in the correlation also increases the forest error rate. A tree with a relatively low error rate is a strong classifier, so an increase in the strength of the individual trees decreases the overall error rate of the forest. There are several notable highlights of random forests. They can run with a high level of efficiency on large databases, as thousands of input variables may be handled without the need for variable deletion. Internal unbiased estimates of the generalization error are generated during the course of forest building. Even when much of the data are missing, a high level of accuracy can be maintained. Furthermore, random forests offer an experimental method for the detection of variable interactions. They are a fast and efficient method that avoids overfitting data.

1.6 References


Holloway, D. T., Kon,DeLisi, Charles "Machine Learning for Regulatory Analysis and Transcription Factor Target Prediction in Yeast."


Chapter 2

Statistical Approaches to Identify Androgen Response Elements

Li Li, Steffen Heber, Qiang Zhang, and Melvin E. Andersen
2.1 Abstract

DNA-binding transcription factors play an integral role in transcription regulation. The annotation of their binding sites in upstream regions of human genes is essential for constructing a genome-wide regulatory network. Transcript factor binding sites (TFBS) in gene promoter regions can be predicted by using computational methods, such as Support Vector Machine (SVM), Hidden Markov Model (HMM), and Random Forests (RF), which summarize sequence patterns of experimentally determined TFBSs. Androgen receptors (AR) are ligand-dependent transcription factors, belonging to the class of steroid hormone receptors. They play an important role in male reproductive functions, and regulate gene transcription by directly binding to androgen response elements (ARE) in target gene promoters. The aim in this study is to use bioinformatics tools to identify and characterize AREs based on sequence information and other sequence-related features. Three statistical methods were explored to strengthen the prediction of putative AREs in the human genome. Instead of using gene expression profiles, we focused on the sequence motifs in the upstream promoters to construct models. The implementation of the HMM search method was built from multiple sequence alignments of known ARE sites. An alternative approach, the SVM model, generated a separating hyperplane between positive and negative ARE classes based on the selected features, including CpG islands, HMM scores, conservation score, and $k$-mer occurrences. Cross-validation results indicated that both models provided good sensitivity and specificity in identifying AREs, with an accuracy of at least 80%. Additionally, random forests (RF) model built multiple decision trees from different subsets of entities responsible for the separation between the classes. Similar to the SVM model, the RF classifier showed low prediction error rate. It is the first time that HMM and SVM have been applied for the prediction of putative AREs within the human genome.

**Key words:** transcription factor binding site, androgen response element, hidden markov model, support vector machine.
2.2 Introduction

Computational methods for modeling and identifying DNA regulatory elements have been developed over the past two and a half decades (GuhaThakurta 2006). In silico prediction of transcription factor binding sites (TFBS) remains important, not only to detect rarely expressed genes but also to analyze genes with unknown function. TFBSs are usually short, around 5-15bp. Potential binding sites can occur very frequently by chance in larger genomes such as the human genome. Hence, it can be difficult to predict TFBS using simple sequence searching tools like BLAST. However, some methods attempt to predict TFBS for a specific transcription factor given a collection of known binding sites already available (Osada 2004; Wang 2006). Our proposed method in this paper is relevant to the latter approach.

Androgen receptors (AR) belong to the steroid hormone receptor superfamily of ligand-inducible transcription factors. It plays key roles in male sexual differentiation and pubertal sexual maturation, and is essential for the maintenance of male reproductive function and behavior through mediating the effect of androgens (Sheckter 1989; Keller 1996; Heinlein 2002). AR contains distinct domains for ligand binding, DNA-binding, and NH$_2$-terminal domains, and is responsible for transcriptional activation, subcellular localization, and dimerization (Lee and Chang 2003). AR usually modulates gene expression by interacting with specific elements in the regulatory regions of various target genes through DNA-binding domains. Upon ligand binding, it undergoes a substantial conformational change that results in recognizing androgen response elements (ARE) in the promoter region of its controlled genes. The DNA-receptor complexes then exert a positive or negative effect on gene transcription by recruiting either coactivators or corepressors to their target genes (Heinlein 2002; McEwan 2004; Chang 2005). The ARE sequence is highly conserved and is composed of two hexanucleotide half-sites separated by a three-nucleotide spacer. Therefore, AR binds as homo- or hetero- dimers to these hexamers, each dimer unit interacting with one of the hexamers (Roche 1992). Variations in the half-site
sequences enable specific binding of a given receptor dimer to multiple target genes with different affinities, resulting in different levels of activation of AR-mediated gene transcription (Schwabe 1993). Identification and characterization of consensus AREs to evaluate their hormonal gene regulation is of importance for the understanding of potential mechanisms of androgen-specific gene transcription and ARE-associated diseases.

Biologists study gene regulation and protein-DNA interaction based on laboratory experiments such as chromatin immunoprecipitation (ChIP) and electrophoretic mobility shift assays (EMSA). These wet-lab experiments allow for the identification of putative AREs; however, they are not only time consuming and labor intensive, but also costly. On the other hand, computational methods can be used to identify TFBS, at the expense of accuracy, but with the benefit of being easier to conduct with available resources (Wagner 1999; GuhaThakurta 2001; Sinha 2003). In general, potential binding sites can be determined by scanning the target sequence with the consensus sequence. This prediction method only works well for highly conserved sequences. Some state-of-the-art statistical methods are employed to characterize the binding preferences of transcription factors, and to identify their putative sites in genomic sequences. For instance, Hidden Markov Models (HMM) (Xu 2005; Abnizova 2006), Support Vector Machines (SVM) (Holloway 2005; Passerini 2006), and Random Forests (RF) (Diaz-Uriarte 2006; Hoffmann 2006) have been found to be effective and robust for classification. Simply speaking, HMM provides a natural way of incorporating both sequence information and probability models of structure into a form suitable for analyzing consensus DNA sequences, while the SVM approach is one of the most popular supervised learning algorithms that determine the maximum-margin hyperplane between two classes of training datasets. The RF model is a collection of tree predictors created by using bootstrap samples of the training data with random feature selection in tree induction (Breiman 2001). Applied to predict cis-regulatory modules and TFBS, all of these methods produced reliable prediction results (Holloway 2005; Grau 2006; Wu 2006).
In this paper we utilized these three statistical approaches with the goal of identifying AREs. Within the prediction model framework, the classifiers were developed based on the biological sequence information acquired from the Genome Browser Database. The models were trained to recognize sequence features and conservation patterns that distinguished between known regulatory regions and nonfunctional sequences. Sensitivity and specificity were then assessed by analyzing the performance of the classifiers. After the evaluation, the classification models could finally be used to predict new, unknown AR binding sites in target genes.

2.3 Methods

2.3.1 Data Collection

We have gathered a collection of experimentally validated AR binding sites from an assortment of biomedical literature. The sites and the upstream sequences have been manually assembled to ensure data consistency. We retained in the data only those genes confirmed by ENSEMBL database (www.ensembl.org). To avoid overfitting, the sites for the same factors in promoters from orthologous genes were discarded. All of the binding sites were from Human. For the negative AREs, we randomly selected upstream sequences of housekeeping genes (Eisenberg 2003) whose sequences contain no known AREs.

2.3.2 Feature Selection

Feature selection is an essential data processing step prior to applying machine learning methods. It is important to select features which are most relevant to AR binding site classification. The features used in this study include conservation score, CpG islands, regulatory potentials (RP), and $k$-mer. Conservation scores show a measurement of evolution conservation in eight vertebrates, including mammals, based on a phylogenetic footprinting approach, which has been taken to identify regulatory elements in the noncoding portion of genomes (Siepel
2005). CpG islands are common near transcription start sites, and associated with promoter regions (Hannenhalli 2001). Five-way RP are computed from alignments of human, chimpanzee, mouse, rat, and dog (Kent, Sugnet et al. 2002). RP scores compare frequencies of short alignment patterns between known regulatory elements and neutral DNA. High RP scores are often combined with conserved TFBSs. All of the above features are intended to be highly discriminative properties for transcription factor binding sites. Their numerical values were obtained through the University of California at Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu/). In addition to the specifically identified ARE site, the DNA sequence flanking ARE is known to be essential for AR-mediated transcriptional activity (Nelson 1999). Therefore, sequence-based features corresponding to all possible DNA sequence variants of a given length \( k \) (\( k \)-mer) in the promoter neighborhood were calculated. We extended each ARE motif sequence upstream by 20bp and downstream by 20bp, and then counted the number of times each 4-mer appeared in both positive and negative sequences. To select the \( k \)-mers specific to AR binding sites, F-scores were calculated to measure the discrimination of two classes (Chen 2005). Given the training vector \( \chi_k, k = 1, \ldots, 4^k \), the number of positive and negative instances are \( n_+ = n_-= 40 \), the F-score of the \( i \) th feature is defined in equation (1).

\[
F(i) = \frac{\left( \bar{x}_i^{(+)} - \bar{x}_i \right)^2 + \left( \bar{x}_i^{(-)} - \bar{x}_i \right)^2}{\frac{1}{n_+ - 1} \sum_{k=1}^{n_+} \left( \chi_{k,j}^{(+)} - \bar{x}_i^{(+)} \right)^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} \left( \chi_{k,j}^{(-)} - \bar{x}_i^{(-)} \right)^2}
\]

\( \bar{x}_i^{(+)} \), \( \bar{x}_i^{(-)} \), \( \bar{x}_i \) are the average of the \( i \) th features for the positive, negative, and whole data sets, respectively. \( \chi_{k,j}^{(+)} \) is the \( i \) th feature of the \( k \) th positive instance, and \( \chi_{k,j}^{(-)} \) is the \( i \) th feature of the \( k \) th negative instance. The retrieval of these features was done with Perl and R scripts.
2.3.3 Construction of Hidden Markov Model

A HMM is a probabilistic model to produce the observed sequences from an underlying model with clusters of binding site motifs. The model (Figure 2.1) consists of a set of states, where each state can emit symbols (nucleotides here) based on a probability distribution. The emission probability for a certain nucleotide is specific for each state. States are connected in a chain-like structure, where the probability of moving from one state to another is termed a transition probability (Eddy 1996). It is convenient to think of a HMM as a model that generates sequences. Under the HMM, the complete sequence likelihood is defined, a product of the transition probability and the emission probability. The HMM algorithm works in a simple way: starting in an initial state, we choose a new state with some transition probability, either staying in state one with transition probability \( t_{1,1} \), or moving to state two with transition probability \( t_{1,2} \); each state-transition generates a residue with an emission probability specific to that state. The fundamental equation relevant for this process is a restatement of Bayes' rule. Let \( \chi_i \) represent state at \( i = 1,2,...,N \) and \( y = y_1 y_2 ... y_N \) represent the path across all possible states. Notice that the path that maximizes \( \Pr(y | \chi) \) must be identical to the path that maximizes \( \Pr(y | \chi) = \frac{\Pr(\chi, y)}{\Pr(\chi)} \). The path through the HMM maximizes \( \Pr(\chi_{i+1}, y_{i+1}) = \max \{ \Pr(\chi_{i+1}, y_{i+1} \mid \chi_i, y_i) \Pr(\chi_i, y_i) \} \) where \( \chi_i \) includes states \( \chi_1 ... \chi_i \), and \( y_i \) is the path \( y_1 ... y_i \) that maximizes the \( \Pr(\chi_i, y_i) \). The HMM model then applies this algorithm to find the most probable path of hidden states that emits the observed nucleotides. Training the ARE HMM requires the careful collection of experimentally verified AR binding sites from the published literature. The consensus sequences of AREs are then used to build the HMM training set by probability-based parameterization following multiple sequence alignment. Given a new sequence, one can match it against the profile HMM to detect potential specific AR target genes. There are several web-based HMM software packages that implement the Forward-Backward, Viterbi and Baum-Welch algorithms (Durbin 1998). Since HMMER (Eddy
profile HMM software package, provides a stable, robust way to capture important information about the degree of conservation at various positions in the sequences, we modify the software so that it could be used for ARE finding.

2.3.4 Support Vector Machine

SVM learning is a popular classification approach that has been applied to several prediction problems in bioinformatics (Passerini 2006; Towsey 2006). Based on statistical learning theory (Vapnik 1998), the central idea of the SVM classification is to find a decision surface that has a maximum distance (margin) from the nearest training data points. The implementation of the SVM approach is as follows. First, map the input vectors to a feature space, possibly in a high dimension, either linearly or non-linearly. Secondly, within the feature space, seek an optimized boundary to separate two classes with the least error and maximum margin (Jiang, Yamauchi et al. 2006). Support vectors are data points that lie on the margins (Figure 2.2). Consider a classification problem with training dataset pairs \( \{(x_i, y_i)\} \) where \( x_i \in \mathbb{R}^d \), and \( y_i \in \{+1, -1\} \) denotes two ARE classes: \( y_i = +1 \) indicates the sample \( i \) being positive ARE and \( y_i = -1 \) indicates the negative class. A linear separating hyperplane generated by the SVM is given by

\[
\{x : f(x) = x^T \beta + \beta_0 = 0\}
\]

where \( \beta \) is a unit vector; \( \|\beta\| = 1 \) and maximal margin \( C = \frac{1}{\|\beta\|} \). A classification rule induced by \( f(x) \) is \( \text{sign}(y_i(x^T \beta + \beta_0)) \), which gives the signed distance from a data point \( x \) to the bounding plane. It is not difficult to find a function

\[
f(x) = x^T \beta + \beta_0 \text{ with } y_i f(x_i) > 0
\]

to separate two classes (Hastie 2001). In short, SVM classification is used to achieve maximum separation between two classes. Feature selection is critical in reducing classifier complexity and cost, and improving model accuracy and generalization. The selected features we mentioned above, along with CpG islands, RP score, interspecies conservation, HMM scores, and \( k \)-mer frequency, are the most commonly
considered ones associated with TFBSs. An R-based SVM package provides a training function with standard and formula interfaces that allow for constructing ARE-SVM classifiers in R modules.

### 2.3.5 Random Forest

A random forest, as its name suggests, is a collection of identically distributed trees. The algorithm works in the following way (Breiman 2001). First, for each iteration in RF, an $n_{\text{tree}}$ bootstrap sample is randomly drawn from the training data. Secondly, a classification tree is induced from the bootstrap sample to maximize size without pruning. At each node, instead of searching through all variables for the optimal split, a tree only searches through a random sample $m_{\text{try}}$ of the variables. Finally, the two steps are repeated for a number of times before the predictions are made by majority vote of the trees for classification. The performance of RF depends on the correlation between trees and the strength of each individual tree. The idea is to maintain the strength of the trees while reducing their correlation with each other. The simulation and analyses were carried out with R, using package randomForest (Liaw 2002).

### 2.3.6 Performance Assessment

An important step in classification is to assess the accuracy of the model performance in a statistically significant way. We used $k$-fold cross-validation to estimate model accuracy. The procedure of $k$-fold cross-validation involved partitioning the training data into $k$ disjoint subsets of approximately equal size. One of the $k$ subsets was used as the test set and the other $k-1$ subsets were put together to form a training set. Then, the average sensitivity and specificity across all $k$ trials were computed. For prediction results, we used the following abbreviations for empirical quantities: TP (# true positives), TN (# true negatives), FP (# false positives), FN (# false negatives), Ac (Accuracy), Sn (sensitivity), and Sp (specificity). Sn, Sp and Ac were calculated according to standard formulae as follows.
Sensitivity is the percentage of correctly predicted real sites among all positives sites and specificity is the percentage of correctly predicted false sites among all negatives sites. Accuracy is the percentage of correctly predicted sites among all sites. The advantage of cross-validation method is that it matters less how the data gets divided. Every data point is placed in a test set exactly once, and gets to be in a training set \( k \)-1 times. The implemention of 10-fold cross validation analysis was carried out in a Perl script. The results of SVM- and RF-based predictions were compared with those achieved with the HMM model.

2.3.7 Genome-wide Identification of Putative AREs

After extracting upstream sequences of human RefSeq, we conducted a genome-wide search for AREs among conserved elements in the -5,000bp upstream of the transcription start site in the GenBank™ annotation of the genome reference sequences (NCBI Genome Build 36 Ver3, March, 2006). Of 23570 genes, the upstream sequences collected above were applied to the HMM model to detect potential AR binding sites.

2.4. Results

2.4.1 HMM Model Architecture

The performance of predictive models is dependent on the quality and availability of training datasets. Here we collected a set of 40 putative AR binding sites from the biomedical literature (Table 2.1). All binding sites were from Human. To create a multiple sequence alignment, these 40 validated AREs were aligned to generate a profile, as shown by the corresponding sequence
logo (Figure 2.3). The logo provided a precise description of the ARE binding site, whose consensus sequence was composed of one canonical half-site of AGAACA and separated by three nucleotides from the other half-site TGTTCT. The overall column height of symbols within the stack indicated the relative frequency of each nucleic acid at that position. The architecture of the HMM model (Sandelin, 2005) was then chosen based on the biological characteristics of these AR binding and the volume of training data (Figure 2.1). From the background state B, it was possible to move to the match state, which consisted of two half-site models separated by a spacer-state classifier. These interacting match states were able to emit all nucleotide symbols with probabilities according to their fitted parameters. However, a match state might be connected to a mismatch state, allowing for the possibility of sequence variability in the model. Transition events were restricted to match, mismatch and spacer states. The last non-interacting match state demanded a transition to the end state. Only the positive ARE training sequences were aligned to construct a positive HMM model through the HMMER package. The negative HMM model was implemented by setting up the probability of all four nucleotides as 0.25. All other parameters in HMMER were set to default values. Given a test sequence categorized by comparing its bit-score produced by the positive HMM to that produced by the negative HMM, the HMM model reported a bit-score and an E-value for every testing sequence. The bit-score indicates how likely the sequence could be a member of positive training sequences, while the E value describes the random background noise that exists for matches between sequences (http://www.ncbi.nlm.nih.gov).

2.4.2 HMM Model Evaluation

The HMM model was validated using the training data set as mentioned in methods and tested in a 10-fold cross-validation. 40 known positive and 40 negative sequences were applied for testing the prediction accuracy of the HMM model. The sensitivity and specificity were calculated for all possible cut-off levels (Figure 2.4A). A cut-off e-value of 0.7, for example, taking a bit score of 2
or higher to predict ARE, represented a reasonable compromise between sensitivity and specificity. Both the sensitivity and specificity were 0.8, supporting a good predictive ability of the HMM model.

Classification accuracy can be illustrated through Receiver Operator Characteristic (ROC) curves, constructed by plotting the sensitivity vs. false positive rate (1 – specificity) for all possible cut-off levels. The area under the ROC curve (AUC) has a number of desirable properties and is becoming increasing popular as a performance measure: the greater the area under the curve, the lower the overall prediction error (Bradley 1997). Calculating AUC is of importance to evaluate the classifier's ability of correctly identifying ARE over the full range of thresholds. We thus utilized the AUC for the model performance measurement and obtained an AUC of 0.73 (Figure 2.4B), indicating a good predictive performance.

2.4.3 SVM Model Evaluation
The first important step in developing a SVM model is to extract relevant features, which can be used as the inputs of SVMs. Based on the positive and negative ARE gene subsets, features are transformed to numerical values through UCSC Tracks (Supplemental Table 2.S). The CpG percentage is the ratio of CpG nucleotide bases (twice the CG count) to the length of sequences (Gardiner-Garden 1987). CpGs are present at significant levels when the score is greater than 50%. For regulatory potential, the default range is from 0 to 0.01. Score values below 0.01 indicate a resemblance to alignment patterns typical of neutral DNA, whereas score values above 0.01 indicate a very marked resemblance to alignment patterns typical of regulatory elements in the training set. Conservations are displayed by pairwise alignments of each species to the human genome, reflecting the phylogeny of species. The larger the score values, the higher levels of overall conservation. HMM scores for positive and negative ARE classes were calculated respectively through HMM model. Sequence-based $k$-mer features were generated
with a set of 256 ($4^4$) substrings representing the promoter motifs. Table 2 shows ten 4-mers as top-ranking classifiers by calculating F-score using equation (1). The SVM model was then trained using these high dimensional features as training vectors. The radial basis (RBF) kernel function was selected to train the SVM, where the RBF kernel function was defined as:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$$

The kernel function parameter $\gamma$ and the parameter $C$, which controls the complexity of the decision function versus the minimum training error, can be determined by running a two dimensional grid-search. This means that the values for pairs of parameters ($C, \gamma$) are generated in a predefined interval with a fixed step. Without knowing beforehand which $C$ and $\gamma$ are the best for one problem, a stratified ten-fold cross-validation procedure is used to evaluate the performance of the SVM classifiers, which depends on the combination of $C$ and $\gamma$ (Huang 2006). The grid-search is easily parallelized because each ($C, \gamma$) is independent. The best performance of our SVM model was given by setting $C$ to 8 and $\gamma$ to 1 (Figure 2.5A). After the best ($C, \gamma$) was determined, the whole training set was used to generate the final classifier. The decision values of binary classification problems were visualized via a plot with a two-dimensional projection of the data with highlighting classes and support vectors in colors (Figure 2.6). To assess the performance of the SVM model, a ROC curve was generated. The area under the ROC curve for the SVM model is 0.8, which indicates good prediction (Figure 2.5B).

2.4.4 Comparison of HMM, SVM, and RF Models

In this study, three major statistical approaches, HMM, SVM and RF have been built using the same original training set to generate an androgen receptor binding site profile. The performance statistics of these models for ARE identification are illustrated in Table 2.3. Overall, all of the three approaches showed good performance; both sensitivity and specificity were at least 0.8. However, the corresponding sensitivity and specificity of the SVM classifier built by the HMM
score were 0.74 and 0.76 respectively. Using top-ranked \( k \)-mer features, the SVM model only gave an accuracy of 0.76, while the RF model predicted better with an accuracy of 0.85. Combining CpG islands, conservations, RPs, and HMM scores together, the SVM model resulted in scores of 0.82 and 0.83 for sensitivity and specificity, respectively. When both the \( k \)-mer features and HMM scores were considered, the performance of the SVM and the RF models were improved giving better prediction than the HMM model. Additionally, we tested how many positive and negative AREs in the training dataset could be detected by these models. It is likely that the HMM approach provides the similar power for prediction, compared to the SVM and RF models.

2.4.5 Genome-wide Identification of AREs

We performed a whole genomic search in the −5,000bp upstream promoter regions of the human RefSeq genes, and the potential target AREs were identified by comparing those genes with the HMM profiles which was the multiple sequence alignment of known AREs. This analysis resulted in around 390 AR binding sites identified. However, the number of sites was larger than the expected frequency in random DNA sequences as calculated by the total number of base pairs in the genome divided by the frequency of a sequence with specified based pairs at 12 positions, which leads to 192 sites (Horie-Inoue 2004). Several known ARE target genes including kallikrein-related peptidase 3 (KLK3), 24-dehydrocholesterol reductase (DHCR24), phytanoyl-CoA 2-hydroxylase (PHYH), were recovered from this analysis.

2.5 Discussion

With the genome sequences of many organisms completed, a full understanding of transcription factors, their DNA binding sites, and their interactions will be the next important issue of genomics (Kim and Kim 2006). The availability of a large number of genome sequence data makes it possible to systematically analyze binding-site patterns in order to provide biological
interpretations. Once a regulatory sequence motif has been identified, the next goal is frequently to identify candidate target genes that may be regulated through it, potentially by a TF that may bind to it, and so statistical and computational methods can be used to detect these sites (Bulyk 2003).

In the present study, using the literature as a guideline, three statistical approaches were employed to build classifiers for predicting human AR target genes. The searches were based on specific models derived from experimentally verified AREs. All of the three computational approaches described the architecture of DNA regulatory elements and were capable of tapping into the tremendous amount of statistical information for classification. The HMM approach was constructed based on sequence context of positive and negative AREs, whereas the SVM encapsulated a significant amount of discriminatory information in the choice of its kernel to yield the optimal separating hyperplane. The RF model collected tree predictors created by using bootstrap samples of the training data and random feature selection in tree induction. The results showed that all of the three methods provided similar performances with overall 80% accuracy rate.

There are a few reports on the prediction of AR-regulated genes (Horie-Inoue 2004; Masuda 2005). These efforts concentrate on screening of palindromic ARE sequences in the human genome. However, a direct search using conserved sequence is not flexible enough to reflect the variability of ARE sequences. We have used a simple approach to build a reliable HMM model compatible with the HMMER package. Several features make the HMM model attractive. First, profile HMM model insertions and deletions and allow fragment matches to the model (Eddy 2003). While insertion and deletions are rare in the functional half sites, they can occur with higher frequency in the spacer regions that are much more divergent (Sinha 2002). Secondly, all hits returned by the HMM model are subject to a bias composition filtering based on a null model
that is computed for each alignment and leads to a rescoring of the hits penalizing the ones for which the nucleotide composition is biased (Marinescu 2005). A binding sequence is then easily represented as a HMM that gives the probability of each nucleotide occurring at a particular position dependent on the nucleotides at preceding positions. By investigating characteristics of training data, we presented an alternative SVM approach for the identification of AREs. One remarkable property of the SVM approach is that their ability to learn can be independent of the dimensionality of the feature space (Joachims 1998). The idea behind the SVM is to construct an optimal separating hyperplane between positive and negative classes. Optimal means that the distances of the hyperplane to the closest points (support vectors) of either class are maximized (Mocellin, Ambrosi et al. 2006). Hence, only these data are necessary to build a SVM classifier, while the remaining training cases do not contribute significantly. Another feature of SVM is that they have the potential to handle a large number of feature spaces. The SVM approach allows for the ability to construct predictive models with a large generalization power, even in the case of a large dimensionality of the data when the number of observations available for training is low (Plewczynski, Tkacz et al. 2005). When SVMs linearly or non-linearly map a high dimensional feature space, different information is taken into account, and more predictable accuracy is achieved. Even features ranked lowest still contain considerable information and are somewhat relevant. A SVM classifier using only the worst features has a performance far superior to random. For RF, there are two significant advantages over other methods. First, there is only one parameter ($m_{t_{ry}}$) to adjust, and the result is usually not sensitive to it. Second, the built-in cross-validation via the use of out-of-bag data gives quite accurate estimates of test set error, and offers quite effective protection against overfitting (Breiman, 2001). These above arguments give theoretical evidence that SVM and RF should perform well in classification (Kim, 2006).

We compared the predictive performance of all of the three statistical models. When limited data sets are handled by these analyses, it is important to perform a leave-one-out test for the whole
available data set to gain a better estimate for the generalization power of analysis than is obtained using only a separate testing set. Thus, ten-fold cross validation is carried out. In contrast to HMM, the SVM model is definitely more encouraging because the sensitivity (0.82) and specificity (0.83) seem favorable. Differences in feature selection among HMM, SVM and RF might contribute to discrepancies between the search methods. The relative poor performance of the reduced SVM model with fewer features is unsurprising because the sequence features extracted are likely to contain other transcription factor binding sites and promoters, not specific to AREs. By removing features that are irrelevant and adding those that are closely correlated, the accuracy of SVM prediction can be improved in future studies. A genome-wide promoter search with the HMM model has identified 390 genes, whose frequency of occurrence is more than the initial expectation of approximately 190 sites (Horie-Inoue 2004).

Although existing database TRANSFAC is widely useful for prediction of TFBSs, the position-specific weight matrix (PWM) for AREs in TRANSFAC was constructed based on only 7 known AR binding sites. Their ARE motif AGWACATNWTTCT could only be recognized when the testing sequence is a good match. On the contrary, our approaches are flexible; allowing the variation in the sequence, such as mismatches and gaps. The ability to detect interactions between AR and its response elements offers the opportunity to study the mechanisms of gene expression mediated by AR. The conserved ARE sequence motifs could possibly be involved in the tissue specificity of transcription by mediating either a positive or negative effect on the basal promoter activity (Fabre, Manin et al. 1994). However, it remains to be elucidated whether the ARE sequence is the major molecular determinant of AR specificity. Identifying direct targets of AR may allow the detection of critical changes of AR-related disease progression (Masuda 2005). For instance, prostate specific antigen (KLK3) recovered by the HMM search is an indicator for the detection of prostate cancer and for monitoring disease progression (Magklara 2002). Therefore, identifying specific AR-regulated genes may contribute to the systematical elucidation
of the gene regulatory network mediated by androgen, which will be pivotal for the development of diseases treatments. Our prediction models will be a valued resource for researchers attempting to identify genes directly regulated by AR transcription factors.

In conclusion, we have developed bioinformatics-based methods for the prediction of candidate AREs in the human genome. AR binding site analyses by the SVM, HMM and RF methods allow for quick and accurate predictions of ARE position in the unknown sequences. The choices of both machine learning algorithms and optimal attribute sets for the selected algorithm are critical for the prediction tasks. Identification of these novel AREs may ultimately lead to the development of better therapies for the treatment or prevention of androgen-related diseases. This study is the first in incorporating AR promoter context in a HMM, SVM and RF topology.

2.6 Acknowledgement

We are grateful for many valuable discussions with Tianyuan Wang and Longlong Yang. This work is supported by funds from the Long-Range Research Initiative of the American Chemistry Council.
2.7 References


Table 2.1. Summary of experimentally verified androgen response element.

<table>
<thead>
<tr>
<th>Positive ARE Gene</th>
<th>ARE sequences</th>
<th>Chromosome Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>F9</td>
<td>AGCTCAGCTTGTTCT</td>
<td>chrX 138440525</td>
</tr>
<tr>
<td>PRKX</td>
<td>AGAACATGTGGTTCT</td>
<td>chrX 3181583</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>AGCACGGCGAGGTCTCC</td>
<td>chr6 36754272</td>
</tr>
<tr>
<td>PCYT1B</td>
<td>AGAACACTATGGTTCT</td>
<td>chrX 23945954</td>
</tr>
<tr>
<td>ACTN1</td>
<td>TGAACCTTCCTTGTTCT</td>
<td>chr14 68517624</td>
</tr>
<tr>
<td>BCOR</td>
<td>AGAACAAAGTTGTTCT</td>
<td>chrX 38970910</td>
</tr>
<tr>
<td>ANKII</td>
<td>AGAACACACTTGTTCT</td>
<td>chr5 14925820</td>
</tr>
<tr>
<td>O60275</td>
<td>AGAACATTATAGTTCT</td>
<td>chrX 52321185</td>
</tr>
<tr>
<td>AZGP1</td>
<td>AGAATAGCAGTTGCTC</td>
<td>chr7 99412366</td>
</tr>
<tr>
<td>PFKFB1</td>
<td>AGAACATAATGTTCT</td>
<td>chrX 53952953</td>
</tr>
<tr>
<td>B2M</td>
<td>AGATCTTAATCCTCT</td>
<td>chr15 42790591</td>
</tr>
<tr>
<td>NM_144657</td>
<td>AGAACATGTGGTTCT</td>
<td>chrX 82472510</td>
</tr>
<tr>
<td>DHCRR24</td>
<td>AGAACATCTATTTCC</td>
<td>chr1 55128366</td>
</tr>
<tr>
<td>NM_024921</td>
<td>AGAACATTCTGTTCT</td>
<td>chrX 83327890</td>
</tr>
<tr>
<td>KLK4</td>
<td>AGGCACTGGTTCT</td>
<td>chr19 56108350</td>
</tr>
<tr>
<td>NDRG1</td>
<td>TGATTAACCTGTCTTCTCT</td>
<td>chr8 134379654</td>
</tr>
<tr>
<td>PHYH</td>
<td>AGAACAGAAGTTTCT</td>
<td>chr10 13383369</td>
</tr>
<tr>
<td>RAB4</td>
<td>ACAAAAGTATGTACT</td>
<td>chr1 227472272</td>
</tr>
<tr>
<td>UGDH</td>
<td>AGACCAGCCTGGTCA</td>
<td>chr4 39206926</td>
</tr>
<tr>
<td>SORD</td>
<td>AAATACCCGTCTCTCT</td>
<td>chr15 43100638</td>
</tr>
<tr>
<td>NJX3A</td>
<td>AGAACATTCTTGCT</td>
<td>chr8 23597605</td>
</tr>
<tr>
<td>PSDR1</td>
<td>AGAACCTGATTCTCA</td>
<td>chr14 67232450</td>
</tr>
<tr>
<td>PIGR</td>
<td>GGTCTTTTCAGTTCT</td>
<td>chr1 205189743</td>
</tr>
<tr>
<td>TMPRSS2</td>
<td>AGCAGATTCTGGTCT</td>
<td>chr21 41802750</td>
</tr>
<tr>
<td>TPDS2</td>
<td>AAAACAGATTTATATAT</td>
<td>chr8 81246919</td>
</tr>
<tr>
<td>UNCI3</td>
<td>AAAAGAATCTGTCTCT</td>
<td>chr9 35151710</td>
</tr>
<tr>
<td>SPAK</td>
<td>TGAACCTGAGTTCTCT</td>
<td>chr2 16812738</td>
</tr>
<tr>
<td>hParl</td>
<td>AGAAGAAATGTACAGG</td>
<td>chr5 76046039</td>
</tr>
<tr>
<td>probasin</td>
<td>ATAGCTCTGTGTTCT</td>
<td>chrX 74106505</td>
</tr>
<tr>
<td>TMEPAI</td>
<td>TGAAGAATGTGGTTCT</td>
<td>chr20 55719263</td>
</tr>
<tr>
<td>SEPP1</td>
<td>TGAACATACTGTCTCT</td>
<td>chr5 42844351</td>
</tr>
<tr>
<td>KLK3</td>
<td>AGAACAGCAGTTCTCT</td>
<td>chr19 34433423</td>
</tr>
<tr>
<td>TRPM8</td>
<td>GGAACCGTGTTCT</td>
<td>chr2 234486965</td>
</tr>
<tr>
<td>SYTL4</td>
<td>AGAACAGATGTGTTCT</td>
<td>chrX 98772894</td>
</tr>
<tr>
<td>NGFRAP1</td>
<td>AGAACAGCTGTGTTCT</td>
<td>chrX 101412450</td>
</tr>
<tr>
<td>CUL4B</td>
<td>AGAACAGCATGTTCT</td>
<td>chrX 118442517</td>
</tr>
</tbody>
</table>
Figure 2.1. Graphical representation of HMM framework for AR binding sites prediction. Arrows denote possible transitions between states. From the background state, two principal transitions are possible: remain in the background state or move to the match state chains of the half-site. The chain of states consists of two half-site models (sequences), separated by a set of spacer states that can generate 0-3 different spacer configurations.

Figure 2.2. General SVM prediction scheme. The training data features are sketched as dots in a two dimensional feature space, which are classified as +1 (open circle) and –1 otherwise (solid circle). The SVM classifier separates the classes by an optimal separating hyperplane \( \mathbf{x}^T \beta + \beta_0 = 0 \) with maximum margin from the hyperplane to the closest point. The best decision surface is determined by only a small set of points (support vectors).
Figure 2.3. A graphical representation of an aligned set of AR binding sites. Sequence conservation, measured in bits of information, is depicted by the height of a stack of letters for each position in the binding sites.

Figure 2.4. Performance of HMM model. A. Plot of sensitivity and specificity vs. E-value based on 10-fold cross-validation. B. Receiver-operating-characteristic curves (ROC) for the prediction model. It shows the proportion of true positives selected by the HMM versus false positives. The performance is shown by the area under the ROC curve.
Figure 2.5. Performance of SVM model. A. 10-fold cross validation classification rate of SVM trained using a different parameter combination, C=4 to 16 and $\gamma$=0.6 to 2.0. The right scale shows the misclassification error rate. The best parameter pair is obtained using C=8 and $\gamma$=1.0. B. The ROC-curve for SVM model. The area under the ROC curve is 0.8, which means good prediction.
Figure 2.6. A contour plot of the fitted decision values for ARE classification. True classes are highlighted through symbol color; predicted class regions are visualized using colored background. P represents positive class while N represents negative class.
Figure 2.7. Performance of SRF model. These two figures showed the prediction error rates for Random forests method. **A.** The plot shows the OOB error rate by mtry. **B.** The plot shows the error rate by trees. The green line means the positive class error rate, which overall is 0.1, whereas the red line represents the negative class error rate, which is 0.2 and the overall OOB error rate, represented by black line, is around 0.15.
Table 2.2. F-score selection of 4-mer motifs in AR-regulated genes.

<table>
<thead>
<tr>
<th>Recognition sequence</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTCT</td>
<td>0.256</td>
</tr>
<tr>
<td>AGAA</td>
<td>0.255</td>
</tr>
<tr>
<td>GTTC</td>
<td>0.233</td>
</tr>
<tr>
<td>GAAC</td>
<td>0.216</td>
</tr>
<tr>
<td>TGTT</td>
<td>0.168</td>
</tr>
<tr>
<td>TTAA</td>
<td>0.147</td>
</tr>
<tr>
<td>GCGC</td>
<td>0.142</td>
</tr>
<tr>
<td>GGCG</td>
<td>0.128</td>
</tr>
<tr>
<td>AAGT</td>
<td>0.124</td>
</tr>
<tr>
<td>ATGT</td>
<td>0.1225</td>
</tr>
<tr>
<td>AGCC</td>
<td>0.1185</td>
</tr>
<tr>
<td>TCTG</td>
<td>0.1184</td>
</tr>
<tr>
<td>TAGC</td>
<td>0.1164</td>
</tr>
<tr>
<td>GGGC</td>
<td>0.1034</td>
</tr>
<tr>
<td>AACA</td>
<td>0.1028</td>
</tr>
<tr>
<td>GCCT</td>
<td>0.0946</td>
</tr>
<tr>
<td>AGTA</td>
<td>0.0929</td>
</tr>
<tr>
<td>CGCC</td>
<td>0.09</td>
</tr>
<tr>
<td>TTTA</td>
<td>0.0852</td>
</tr>
<tr>
<td>TCTA</td>
<td>0.083</td>
</tr>
</tbody>
</table>
Table 2.3. Summary of the performance for the HMM, SVM, and RF models.

<table>
<thead>
<tr>
<th></th>
<th>HMM</th>
<th>SVM</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k-mer</td>
<td>All features</td>
<td>k-mer</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.80</td>
<td>0.75</td>
<td>0.82</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.80</td>
<td>0.88</td>
<td>0.83</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.80</td>
<td>0.76</td>
<td>0.81</td>
</tr>
<tr>
<td>Positive Prediction</td>
<td>35/40</td>
<td>34/40</td>
<td>35/40</td>
</tr>
<tr>
<td>Negative Prediction</td>
<td>37/40</td>
<td>29/40</td>
<td>38/40</td>
</tr>
</tbody>
</table>
### Supplemental Table 2.S. Genes and their corresponding features.

<table>
<thead>
<tr>
<th>Gene</th>
<th>CpG island</th>
<th>Regulatory Potential</th>
<th>HMM Score</th>
<th>Conservation</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>F9</td>
<td>0.5</td>
<td>0.477404</td>
<td>0.21</td>
<td>0.992126</td>
<td>positive</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>0.81</td>
<td>0.153182</td>
<td>4.4</td>
<td>0.00703937</td>
<td>positive</td>
</tr>
<tr>
<td>ACTN1</td>
<td>0.5</td>
<td>0.319203</td>
<td>0.13</td>
<td>0.0393701</td>
<td>positive</td>
</tr>
<tr>
<td>ANKH</td>
<td>0.81</td>
<td>0.0599002</td>
<td>0.77</td>
<td>0.00255118</td>
<td>positive</td>
</tr>
<tr>
<td>AZGP1</td>
<td>0.5</td>
<td>0.03389</td>
<td>0.67</td>
<td>0.0217323</td>
<td>positive</td>
</tr>
<tr>
<td>B2M</td>
<td>0.79</td>
<td>0.0137741</td>
<td>0.12</td>
<td>0.00775591</td>
<td>positive</td>
</tr>
<tr>
<td>DHCR24</td>
<td>0.5</td>
<td>0.0352366</td>
<td>0.024</td>
<td>0.11674</td>
<td>positive</td>
</tr>
<tr>
<td>KLK4</td>
<td>0.5</td>
<td>0.0352366</td>
<td>0.34</td>
<td>0.11674</td>
<td>positive</td>
</tr>
<tr>
<td>NDRG1</td>
<td>0.5</td>
<td>0.0352366</td>
<td>0.41</td>
<td>0.11674</td>
<td>positive</td>
</tr>
<tr>
<td>PHYH</td>
<td>0.83</td>
<td>0.159765</td>
<td>0.14</td>
<td>0</td>
<td>positive</td>
</tr>
<tr>
<td>RAB4A</td>
<td>1.01</td>
<td>0.0282086</td>
<td>2</td>
<td>0.0105197</td>
<td>positive</td>
</tr>
<tr>
<td>UGDH</td>
<td>0.8</td>
<td>0</td>
<td>0.71</td>
<td>0.00692126</td>
<td>positive</td>
</tr>
<tr>
<td>SORD</td>
<td>0.69</td>
<td>0.21824</td>
<td>1.1</td>
<td>0.000937008</td>
<td>positive</td>
</tr>
<tr>
<td>NJX3A</td>
<td>0.93</td>
<td>0.066002</td>
<td>1.5</td>
<td>0.00143307</td>
<td>positive</td>
</tr>
<tr>
<td>PSDR1</td>
<td>0.5</td>
<td>0.066002</td>
<td>0.55</td>
<td>0.00143307</td>
<td>positive</td>
</tr>
<tr>
<td>PIGR</td>
<td>0.5</td>
<td>0.0324784</td>
<td>0.5</td>
<td>0.00494488</td>
<td>positive</td>
</tr>
<tr>
<td>TMPRSS52</td>
<td>0.73</td>
<td>0.0795879</td>
<td>0.43</td>
<td>0.00761417</td>
<td>positive</td>
</tr>
<tr>
<td>TPD52</td>
<td>1.02</td>
<td>0</td>
<td>2.4</td>
<td>0.992126</td>
<td>positive</td>
</tr>
<tr>
<td>UNC13</td>
<td>0.94</td>
<td>0.181102</td>
<td>0.13</td>
<td>0.456693</td>
<td>positive</td>
</tr>
<tr>
<td>SPAK</td>
<td>0.98</td>
<td>0.00378338</td>
<td>0.083</td>
<td>0.0326299</td>
<td>positive</td>
</tr>
<tr>
<td>MLCB</td>
<td>0.82</td>
<td>0.146077</td>
<td>1</td>
<td>0.00787402</td>
<td>positive</td>
</tr>
<tr>
<td>MVDP</td>
<td>0.5</td>
<td>0.00428324</td>
<td>0.083</td>
<td>0.262094</td>
<td>positive</td>
</tr>
<tr>
<td>C3</td>
<td>0.5</td>
<td>0.00376416</td>
<td>0.057</td>
<td>0.15748</td>
<td>positive</td>
</tr>
<tr>
<td>probasin</td>
<td>0.5</td>
<td>0.00376416</td>
<td>0.17</td>
<td>0.19685</td>
<td>positive</td>
</tr>
<tr>
<td>ODC</td>
<td>0.5</td>
<td>0</td>
<td>0.24</td>
<td>0.0472441</td>
<td>positive</td>
</tr>
<tr>
<td>Slc</td>
<td>1</td>
<td>0</td>
<td>0.089</td>
<td>0.165354</td>
<td>positive</td>
</tr>
<tr>
<td>crp2</td>
<td>0.5</td>
<td>0</td>
<td>0.57</td>
<td>0</td>
<td>positive</td>
</tr>
<tr>
<td>KLK3</td>
<td>0.5</td>
<td>0</td>
<td>0.19</td>
<td>0.0866142</td>
<td>positive</td>
</tr>
<tr>
<td>TRPM8</td>
<td>0.99</td>
<td>0</td>
<td>0.2</td>
<td>0.102362</td>
<td>positive</td>
</tr>
<tr>
<td>PRKX</td>
<td>0.5</td>
<td>0</td>
<td>0.0094</td>
<td>0.0062126</td>
<td>positive</td>
</tr>
<tr>
<td>Gene</td>
<td>CpG island</td>
<td>Regulatory Potential</td>
<td>HMM Score</td>
<td>Conservation</td>
<td>Category</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>----------------------</td>
<td>-----------</td>
<td>--------------</td>
<td>----------</td>
</tr>
<tr>
<td>BCOR</td>
<td>0.8</td>
<td>0.00653606</td>
<td>0.04</td>
<td>0.307087</td>
<td>positive</td>
</tr>
<tr>
<td>PFKFB1</td>
<td>0.97</td>
<td>0.107714</td>
<td>0.07</td>
<td>0.0078740</td>
<td>positive</td>
</tr>
<tr>
<td>NM_144657</td>
<td>0.5</td>
<td>0</td>
<td>2.8</td>
<td>0.0845669</td>
<td>positive</td>
</tr>
<tr>
<td>NM_024921</td>
<td>0.84</td>
<td>0.0102373</td>
<td>0.0013</td>
<td>0.015685</td>
<td>positive</td>
</tr>
<tr>
<td>SYTL4</td>
<td>0.5</td>
<td>0</td>
<td>0.0044</td>
<td>0.0314961</td>
<td>positive</td>
</tr>
<tr>
<td>NGFRAP1</td>
<td>0.88</td>
<td>0.00870768</td>
<td>0.0012</td>
<td>0.075748</td>
<td>positive</td>
</tr>
<tr>
<td>CUL4B</td>
<td>0.78</td>
<td>0.830472</td>
<td>0.0111</td>
<td>0.003981</td>
<td>positive</td>
</tr>
<tr>
<td>ACTG1</td>
<td>0.5</td>
<td>0.0845814</td>
<td>7.5</td>
<td>0.00787402</td>
<td>negative</td>
</tr>
<tr>
<td>FLNB</td>
<td>0.5</td>
<td>0</td>
<td>5.7</td>
<td>0.00787402</td>
<td>negative</td>
</tr>
<tr>
<td>ACTB</td>
<td>0.5</td>
<td>0.174097</td>
<td>6.5</td>
<td>0.755906</td>
<td>negative</td>
</tr>
<tr>
<td>ALDOA</td>
<td>0.5</td>
<td>0</td>
<td>0.18</td>
<td>0</td>
<td>negative</td>
</tr>
<tr>
<td>GAPD</td>
<td>0.5</td>
<td>0.115003</td>
<td>1.7</td>
<td>0.00787402</td>
<td>negative</td>
</tr>
<tr>
<td>PGK1</td>
<td>0.5</td>
<td>0</td>
<td>0.8</td>
<td>0</td>
<td>negative</td>
</tr>
<tr>
<td>RPS27A</td>
<td>0.5</td>
<td>0.038181</td>
<td>3.7</td>
<td>0.992126</td>
<td>negative</td>
</tr>
<tr>
<td>RPL19</td>
<td>0.5</td>
<td>0</td>
<td>2.6</td>
<td>0.00307874</td>
<td>negative</td>
</tr>
<tr>
<td>RPL11</td>
<td>0.5</td>
<td>0</td>
<td>5.6</td>
<td>0.00150394</td>
<td>negative</td>
</tr>
<tr>
<td>NONO</td>
<td>0.5</td>
<td>0.0781072</td>
<td>0.48</td>
<td>0.0393701</td>
<td>negative</td>
</tr>
<tr>
<td>ARHGEDIA</td>
<td>1</td>
<td>0</td>
<td>7.9</td>
<td>0</td>
<td>negative</td>
</tr>
<tr>
<td>RPL32</td>
<td>0.5</td>
<td>0</td>
<td>5.5</td>
<td>0</td>
<td>negative</td>
</tr>
<tr>
<td>RPL322</td>
<td>0.5</td>
<td>0</td>
<td>2.6</td>
<td>0.00756693</td>
<td>negative</td>
</tr>
<tr>
<td>RPS18</td>
<td>0.84</td>
<td>0.0183972</td>
<td>5.1</td>
<td>0.015748</td>
<td>negative</td>
</tr>
<tr>
<td>HS PCB</td>
<td>0.73</td>
<td>0.0586135</td>
<td>8</td>
<td>0.005</td>
<td>negative</td>
</tr>
<tr>
<td>LDHA</td>
<td>0.88</td>
<td>0</td>
<td>0.3</td>
<td>0.00748031</td>
<td>negative</td>
</tr>
<tr>
<td>ACTC</td>
<td>0.5</td>
<td>0.0290162</td>
<td>2.6</td>
<td>0.0314961</td>
<td>negative</td>
</tr>
<tr>
<td>FLNA</td>
<td>0.5</td>
<td>0.0475113</td>
<td>6.9</td>
<td>0</td>
<td>negative</td>
</tr>
<tr>
<td>FCGBP</td>
<td>0.5</td>
<td>0.0191495</td>
<td>6.4</td>
<td>0.00786614</td>
<td>negative</td>
</tr>
<tr>
<td>FCGR3A</td>
<td>0.5</td>
<td>0.00986701</td>
<td>5.4</td>
<td>0.00787402</td>
<td>negative</td>
</tr>
<tr>
<td>FCGR2B</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>0.0781102</td>
<td>negative</td>
</tr>
<tr>
<td>FCGBP</td>
<td>0.84</td>
<td>0.00711857</td>
<td>2.5</td>
<td>0</td>
<td>negative</td>
</tr>
<tr>
<td>ACTIN</td>
<td>0.50.5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>negative</td>
</tr>
<tr>
<td>NONO2</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>0.0551181</td>
<td>negative</td>
</tr>
</tbody>
</table>
Chapter 3

Androgen Receptor-dependent Gene Regulatory Networks

Li Li, Steffen Heber, Qiang Zhang, and Melvin E. Andersen
3.1 Abstract

Androgens are essential steroid hormones for the physiological maintenance and development of primary and secondary sexual characteristics in males. Understanding the underlying mechanism and dynamics of androgen receptor (AR) regulated gene networks requires knowing the direct target genes whose expression levels are modulated by the androgen signaling pathway. Inspired by the study of transcription network in Arabidopsis, we performed a systematic analysis of promoter regions in the human genome for the occurrences of regular expression patterns with the goal of identifying potential AR regulatory elements. First, we analyzed the gene expression profile of human LNCaP cell line in the absence and presence of exogenous androgen, resulting in over 745 differentially expressed genes. To identify regulatory elements that control androgen-mediated gene expression patterns, sequence-based and motif-based analyses in the upstream regions of co-expressed gene groups were performed. The sequence-based feature selection involves calculating the distribution of nucleotide $k$-mer counts, while the motif-based feature generation method scans the promoter regions to collect the predefined transcription factor binding site (TFBS) motifs for each gene. Several highest ranking $k$-mer motifs in DHT regulated genes had perfect matches to the top-ranked TFBS motif elements, such as GATA, CREB and C/EBP. The list of direct targets containing the above motifs could be used as the basis for complex pathways and sequence analyses, which suggests several interactions between the AR and other transcription factors. Additionally, the distributions of all $k$-mers of length 4 or 5 and TFBSs in the gene’s promoter region were examined as a means to predict AR binding. The binding sites were predicted with a Random Forests approach, which classifies androgen responsive genes with a high sensitivity and an accuracy of almost 70%. The sequence analysis of the functional targets in conjunction with the transcription network enabled us to suggest interactions between AR and other transcription factors in specific target genes.

Keywords: Transcription networks, Androgen Receptor, Random forests
3.2 Introduction

With the completion of the human genomic sequence and the development of high-throughput experimental technology, identifying the transcription factor binding sites (TFBS) has become an important approach to understanding the transcriptional regulatory mechanism. DNA microarray technology, for example, provides useful tools to simultaneously study the expression profiles of thousands of distinct genes in different cell and tissue samples. To systematically understand the biological functional roles of the genes in a genome, the expression profiles for a series of experimental conditions can be grouped into clusters based on similar expression patterns (Li 2006). Each cluster contains co-expressed genes that are more likely to have their promoter regions bound by a common transcription factor (TF). However, co-expression is not co-regulation, and hence it is unlikely that all upstream regions of a set of co-expressed genes are controlled by a single transcription factor and contain the associated regulatory elements (Filkov 2005). Co-expressed genes may contain a large number of motifs that are significantly overrepresented in comparison to random gene sets (Holloway 2006). These motifs probably have important functions, and their corresponding TFs may play important roles in the regulation of genes with a particular expression pattern. Identifying the occurrences of TFBS in the co-expressed gene sequence set may help understand how the regulatory networks govern processes that control signaling and metabolic pathways. Identifying TFBS may also aid understanding other cell processes in response to various internal and external conditions.

The androgen receptor (AR) is a transcription factor that plays a critical role in normal sexual development, metabolism, and maintenance of hormonal homeostasis in the male reproductive system. Interruption of these processes by pharmacological androgen antagonists or environmental endocrine disruptors may cause a range of phenotypic abnormalities, such as prostate cancer, one of the most common life-threatening diseases in men. AR contains distinct functional domains that interact with androgens and bind to regulatory regions in DNA. Upon
ligand binding, it undergoes a substantial conformational change that allows for the recognition of androgen response elements (ARE) in the promoter region. The DNA-receptor complex then exerts a positive or negative effect on gene transcription by recruiting either coactivators or corepressors to their target genes (Heinlein 2002; McEwan 2004; Chang 2005). The mechanisms of action of the ligand-bound AR include not only the ability to activate or repress gene transcription, but also interaction with other transcription regulators. That is, the target genes activated by AR are likely dependent upon other transcription factors present on the gene promoters or enhancers (Phuc Le 2005). Specifically, protein kinase A (PKA) activation could stimulate prostate-specific antigen (PSA) promoter-driven transcription through cAMP responsive element-binding protein (CREBP) in the presence of androgen, where the putative cAMP responsive element is located in the 5’ upstream regulatory region of the PSA gene (Kim 2005). Therefore, understanding the nature of AR action requires knowing not only the set of genes bound and regulated by AR, but also the transcription factors that may interact with the AR.

Since the binding affinities between TFs and their corresponding binding sites are largely affected by the consensus sequence pattern of the TFBS and enhancers (Zheng 2003), analyses of promoter regions of co-expressed genes may provide clues to where and which regulatory sequences may be located. This is the most fundamental information that may help in understanding the molecular basis of biological processes. Several groups have developed diverse algorithms that combine different sources of data to predict transcriptional regulatory mechanisms in the low eukaryotic genomes. Bar-Joseph et al., and Gao et al. combined binding data with gene expression data to identify regulatory networks (Bar-Joseph, Gerber et al. 2003; Gao 2004). Accordingly, the usual computational method to discover regulatory networks is to analyze the promoter regions of these gene groups in order to detect over-represented nucleotide patterns. Pattern discovery methods in the upstream region basically fall into two groups: sequence-based and motif-based. A sequence-based approach includes matched, mismatched
and gapped $k$-mer counts, whereas a motif-based approach works by enumerating or searching a predefined pattern class to find patterns and their frequencies. Perhaps one of the best qualitative descriptions of pattern discovery has come out of the work of Li & Lee on Arabidopsis (Li 2006). We are interested in elucidating the transcriptional regulatory mechanism of androgen-regulated gene expression in humans. However, finding regulatory modules in higher eukaryote genomes remains a challenge because of genome complexity (Kim 2006). Inspired by the study of transcription networks in Arabidopsis (Li 2006), our approach combines both sequence- and motif-based data in a supervised learning scheme to identify androgen responsive genes by searching for upstream patterns common to most genes in the co-expressed gene group. The distributions of $k$-mer counts in gene promoter regions were used as features to differentiate AR targets and non-targets. To identify TFBS motif combinations that control androgen-regulated gene expression patterns, Toucan 2, an application for the discovery of significant cis-regulatory elements, was applied to identify all the genes in human containing each motif in their promoters. After comparing and matching top-ranking $k$-mers with top-ranked regulatory elements, the application of machine learning algorithms for promoter sequence analysis could establish models of transcriptional pathways regulating androgen-mediated gene expression.

3.3 Methods

3.3.1 Microarray Data Analysis

In order to understand the ability of ARs to regulate gene expression, human LNCaP cells, androgen-sensitive human prostate carcinoma cells derived from lymph node, were treated for 0, 6 and 24 hours with dihydrotestosterone (DHT). Three biological replicates were used for each treated condition. After cells were harvested at 6 and 24 h, mRNA was isolated and hybridizations were performed on the Affymetrix Hu.133AB platform (Kazmin 2006). Raw Affymetrix .CEL files obtained from Gene Expression Omnibus (GEO) were loaded into GeneSpring 7.2 software (Agilent Technologies, Santa Clara, CA). The analysis was first performed for per-chip and per-
gene normalization using probe-level data. Due to inconsistent sample preparation, per-chip normalizations were performed to control chip-wide variation in intensity. This procedure uses the mean intensities across all chips as a common baseline, and then fits the data from each chip to this baseline. Usually, 50% percentile normalization was selected to adjust all samples to a certain mean intensity value on each chip. Furthermore, per-gene normalization is necessary to compare the expression profiles of genes that may be expressed at very different levels. The normalization process for each gene was performed by normalizing to the median for replicates. To determine which transcripts were significantly regulated by DHT at 6, or 24 hours, pairwise comparison analyses were performed using time point zero as control, which served as the baseline data. The change in each gene expression was then calculated by determining the fold change (ratio) of the mean intensity of each group. Given two sample groups, for each spot of each gene we computed fold-changes for all pairs of samples derived from the two groups to be compared. The median of these spot fold-changes is used as the overall estimate for the gene fold-changes. The genes with expression ratios of DHT and control of more than 1.6 fold or less than 1.6 fold at one time-course point or more were defined as DHT-inducible genes and DHT-repressible genes, respectively.

3.3.2 k-mer Distributions

It is thought that genes with similar expression profiles or genes with similar functions are likely to be regulated via the same mechanism. Inference of such co-regulated genes suggests that they may contain a binding site for a particular TF or other common regulatory elements. Therefore, those genes may share certain attributes, or features, that other genes do not share, and it is on the basis of particular TFBSs. Here, we use the up-regulated and down-regulated gene classes derived from microarray analysis to generate attributes, such as the number of occurrences of a particular nucleotide sequence of length $k$. The basic operation for a $k$-mer or $k$ nucleotide analysis is the computation of the frequencies for all $k$-mers for a given sequence. All upstream
sequences of differentially expressed genes were collected by downloading the upstream1000 data set from ENCODE (www.genome.ucsc.edu). These sequences were then scanned using a sliding window with size $k$ to generate $k$-mer sub-sequences. For a dataset, the number of occurrences of each of the $4^5$ possible nucleotide sequences of length 5 (5-mers) might be represented by a 1024-component vector, each component of which is the number of times the corresponding 5-mer occurs upstream from one of the genes in the set. Four types of $k$-mers were employed here. 4-mers and 5-mers are defined as exact nucleotide matches of length 4 or 5, respectively. Mismatched 4-mers are defined as 4-mers with one mismatch at position 2, 3, 4 or 5, e.g. ANAAC, AGNAC, AGANC, AGAAN (N represents {A, T, C, G}). Mismatched 5-mers allow one mismatch as well. Both mismatched 4-mers and 5-mers could be used to find motifs with less conserved cores. Only 4-mers and 5-mers were chosen because 6-mers may be too sensitive. Finally, a target sequence was transformed into a high dimensional numerical space of $k$-mers. The occurrence site for each $k$-mer on the entire region of a target sequence was recorded in a hash table. To test the occurrence of any $k$-mer in a hash table, a binary variable, $\chi_n$ was defined as follows:

$$\chi_n = \begin{cases} 
1, & \text{when the } k\text{-mer is found in a target sequence} \\
0, & \text{otherwise} 
\end{cases}$$

All possible $k$-mers and the counts of the occurrence of each gene within a set of promoters were generated using Perl scripts. Finally, the $k$-mer distribution built from the counts of all $k$-mers of length 4, 5, mismatched 4-mer, and mismatched 5-mer in each gene’s promoter region could be used as an input vector to predict whether any given gene has a distribution similar to known targets.

3.3.3 Regulatory Sequence Analysis

The availability of genomic sequence and microarray technologies makes it possible to carry out promoter analysis for co-expressed genes. Since most regulatory elements are in non-coding
regions and show considerable variation in sequence even for the same TF, it is not easy to recognize them. However, binding sites are often preserved through evolution, and thus become apparent. A study of the promoter region may help to understand gene co-regulation at the transcriptional level. Toucan 2 is a web-interface for regulatory sequence analysis on metazoan genomes, especially useful for detecting significant transcription factor binding sites (Aerts 2005). It can be used to screen DNA sequences with a precompiled motif model. The source of motifs is the TRANSFAC 7.0 database. While our target sequences are promoters of co-expressed genes obtained from a microarray experiment, promoters of genes with invariant expression within the same experiment could be used as a good control set, since they are most likely not to contain motifs of interest. After we remotely load upstream sequences of response genes regulated by DHT to Toucan 2, all conserved sequence blocks within 1 kb upstream of the transcription start site can be detected and scored with all position weight matrices of TRANSFAC using MotifScanner. A high threshold would prevent some noise signals from being included with the computational result, while a low threshold could yield more transcription factor binding sites. Applied to the default parameter values and lower thresholds, Toucan 2 provides multiple TFBS sets for each gene in both positive and negative classes. TFBS are similar to \( k \)-mers in that they both use the same training datasets. For instance, the motif profile vector consists of a binary number, with 1 indicating the presence of TFBS and 0 indicating its absence. Therefore, the data for any particular TF consists of a number of different feature vectors with hundreds of dimensions, where each vector represents a gene in the training dataset.

3.3.4 F-score Selection

Variable and feature selection have become the focus of much research in areas of prediction for which datasets with tens or hundreds of thousands of variables are available. To find the best possible combination of features to discriminate the DHT-mediated gene group and non-response group, a relatively simple and fast feature selection approach based on F-scores was presented
by Chen and Lin (Chen 2005). Given the training vector \( \chi_k, k=1, \ldots, 4^k \), the number of positive and negative instances, the F-score of the \( i \)th feature is defined in equation (1).

\[
F(i) = \frac{1}{n_+ - 1} \sum_{k=1}^{n_+} (\chi_{k,j}^{(+)} - \bar{\chi}_i^{(+)})^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} (\chi_{k,j}^{(-)} - \bar{\chi}_i^{(-)})^2
\]

\[(1)\]

Where

\( F(i) \) : F-score value for \( i \)th feature 
\( \chi_k, k=1, \ldots, n \) : Set of \( n \) training vectors 
\( n_+, n_- \) : Number of positive and negative training instances 
\( \bar{\chi}_i^{(+)} \), \( \bar{\chi}_i^{(-)} \) : The average of the \( i \)th feature of the positive, negative, and whole data sets, respectively. 
\( \chi_{k,j}^{(+)} \), \( \chi_{k,j}^{(-)} \) : The \( i \)th feature of the \( k \)th positive and negative instance 

The numerator of the equation indicates the discrimination between the positive and negative sets, and the denominator indicates the discrimination within each of the two sets. To retrieve the top-ranking \( k \)-mers and TFBS motifs responsible for separation between response and non-response classes, Perl and R scripts were written for F-score calculation.

### 3.3.5 Random Forest Prediction

Prediction is performed by transforming sequences into high-dimensional feature representations followed by classification into either negative or positive classes. Our learning task integrates two qualitatively different data sources: motif profile data from regulatory sequences and \( k \)-mer distributions. The optimum subset of features selected through F-score calculation, such as top-ranking \( k \)-mers and TFBS, could best discriminate between positive and negative classes in the training dataset. By using a sliding window with \( k \)-mer width, the distribution of the counts of all \( k \)-mers of length 4 and 5, with or without a mismatch in the gene’s promoter region were examined as a means to predict TF binding (Holloway, 2005). Random forests (RF) have been applied successfully to many classification problems in computational biology (Breiman 2001). A random forest, as its name suggests, is a collection of identically distributed trees (Figure 3.1).
The algorithm works in the following way; from the original sample, two-thirds of the data are used to randomly draw an \( n_{\text{tree}} \) bootstrap sample with replacement. Another one-third is preserved for Out of Bag (OOB) to estimate the classification error. At each node, instead of searching through all variables (features) for the optimal split, a tree only searches through a random sample \( m_{\text{try}} \) of the variables. Finally, the two steps are repeated for a number of times before predictions are made by majority vote of the trees for any test case. Proper AR classifier construction requires that the components of the feature vector comprise \( k \)-mer counts and TFBS motif conservations. The positive and negative training sets are identified from microarray experiments, with each sample represented by a feature vector of attributes. Given a set of true ARE positives and negatives, the RF classifier can classify a testing gene sequence as an AR target or non-target based on the relevant features.

3.3.6 Functional Annotation

To determine whether there is an association between the classes of regulated genes and specific biological responses, we performed a gene ontology analysis. FatiGO is the pioneering tool to extract relevant Gene Ontology (GO) terms for a group of genes with respect to a set of genes of reference (Al-Shahrour 2005). FatiGO takes a list of genes and converts them into a list of GO terms using the corresponding gene-GO association table.

3.4 Results

3.4.1 Differentially Expressed Genes

The microarray data were analyzed considering only fold changes for the identification of differentially expressed genes. Up-regulated and down-regulated genes are those that have been significantly expressed in the treated samples with fold changes greater than 1.6 in the respective direction. Intensity ratios for each gene at each time point (6h, 24h) were compared with the control group accordingly. 745 genes were shown differentially expressed with observed fold
changes greater than 1.6 or less than 0.6 at both time points. After 6h of androgen stimulation, 528 genes exhibited a fold change greater than 1.6-fold. Among them, 337 genes increased more than 1.6-fold while 191 genes decreased more than 1.6-fold. After 24h of androgen exposure, 517 genes exhibited a fold change greater than 1.6-fold. Of these, 433 transcripts increased 1.6-fold, and 84 transcripts decreased 1.6-fold. The scatter plot (Figure 3.2) showed that those genes above the diagonal were considered to be induced by DHT exposure while genes below the diagonal were considered to be suppressed.

3.4.2 $k$-mer Feature Selection

For both the $k$-mer count and the TFBS-based data set as described in the methods, feature selection has been performed in order to achieve the best possible prediction results. Hence, F-scores have been calculated based on equation (1). All upstream sequences were decomposed into a vector of $k$-mers of length 4 or 5, resulting in 256 elements for 4-mers, 1024 elements for 5-mers, 1024 elements for mismatched 4-mers, and 5120 elements for mismatched 5-mers. Figure 3.2 presents the distribution of F-scores against mismatched 5-mer features. The larger the F-score was, the more likely this feature was more discriminative between positive and negative groups. The feature with the highest F-scores was assigned the lowest feature number during sorting. Features with F-scores below the horizontal line were not included into the training data for further analysis. Accordingly, the top-ranked $k$-mer features based on their importance along with the corresponding F-scores are shown in Table 3.1. The selected $k$-mer features were then regarded as the template for decision making, and these templates were chosen as representatives of the training $k$-mers for prediction.

3.4.3 Regulatory Sequence Analysis

In order to perform a thorough regulatory sequence analysis, Toucan 2 was used to find over-
represented TFBS in the proximal promoters of a set of co-expressed genes identified by microarray analysis. For each gene, the occurrences of predefined TFBS motifs were found using a sliding window within the required search region. It searched input upstream sequences and created a graphical representation of the results (Figure 3.3). Toucan 2 reported all motif hits that scored above a threshold set by users. Each input gene is displayed as a straight line whose length is proportional to the length of the original sequence in nucleotides. Individual sites are represented as colored rectangular blocks. The width of the rectangle is proportional to the length of the TF binding site.

Similar to \( k \)-mer selection, the most discriminative TFBS features between response and non-response groups were determined using F-score calculation (Figure 3.4). Table 3.2 gave an overview of the top-ranked motif features along with the corresponding sequences. Features were sorted according to the F-score values calculated.

### 3.4.4 Prediction Results on Training Data Set

By examining datasets for both \( k \)-mer counts and motif detection, the most representative \( k \)-mers and TFBS were determined for the differences between positive and negative groups. Selecting an optimal number of features for classification was a complicated task. Nevertheless, preliminary feature selection using simple F-score calculation was used. The 745 responsive gene set and 745 non-responsive gene set were employed to distinguish between genes regulated by DHT and control, and then used for random forests classification. Based on the out of bag cases for each fitted tree, the random forests approach returned a measure of the OOB error, which was used to assess the effects of parameters. To evaluate the impact of different numbers of top-ranked features on the classification performance of RF, we considered a heuristic by selecting the top-ranked features (from the top 2 through the top 100) for RF prediction. To test the performance of predictors, we computed the following performance
measures: OOB, accuracy, sensitivity and specificity. The results of the best two predictors achieved for different $k$-mer models were reported in Table 3. The mismatched 5-mer models showed a little more improvement than the other three models (4-mer, 5-mer, and mismatched 4-mer), with an accuracy rate of 0.65 vs. 0.60. However, considering TFBS features only, the prediction classifier gave a 0.64 accuracy rate (Table 3.4). When combining two different features together, RF could achieve around a 0.70 accuracy rate with sensitivity of 0.70 shown in Table 3.4 as well.

Several highest-ranking mismatched 5-mer motifs in DHT-regulated genes had perfect matches to top-ranked TFBS motif elements: TAGCGN perfectly matched the M00114-V$\$TAXCREB_01 motif, GCGCNC matched M00056-V$\$MYOGNF1_01, and CACNTG matched M00073-V$\$DELTAEF1_01. Other promoters that have been characterized in AR-mediated gene expression are those for C/EBP, PAX, and GATA factors. The matched regulatory elements between $k$-mers and TFBS are depicted in red (Table 3.5). Out of 745 genes, there are around 300 genes with CREB motif regions, and 240 genes containing GATA motifs. For instance, TMEPAI and SGK3 containing both CREB and AR binding sites are involved in cell growth inhibitory function. It has been demonstrated that TMEPAI, located on chromosome 20q13, exhibits a high level of expression in the prostate (Xu 2000). BRCA1 and IL6R contain both GATA and AR binding sites, causing ligand-dependent and synergistic activation of the AR. A recent study reported that increased serum IL6 was associated with androgen-independent and metastatic prostate cancer progression (Hammacher 2005). Genes with either CREB or GATA binding sites play important roles in primary and cellular metabolic processes. Surprisingly, the C/EBP and PAX4 motifs are enriched in all of the AR-regulated genes. After applying GO annotation, the majority of the genes with C/EBP motifs were found to be involved in protein, ion, and nucleotide binding, and hydrolase activity. Genes that belong to the PAX4 or C/EBP family of transcription factors may be involved in primary or cellular metabolic processes.
3.5 Discussion

The recent accumulation of microarray data has dramatically expedited comprehensive understanding of gene expression profiles. Gene expression profiles with high-dimensional genomic datasets may contain hundreds or thousands of features. Understanding how the expression of thousands of genes is regulated in cells or tissues remains a significant and difficult challenge in biology (Li 2003). Therefore, the focus of research is now shifting from the collection of data to the prediction of functional relationships within specific biological processes aimed at the elucidation of functional gene networks and protein complexes (Werner 2006). In this study, we have pursued the existence of common regulatory motifs in AR-regulated genes identified by microarray analysis. The corresponding promoter sequences were investigated to identify transcription factors that may contribute to specific features of AR-mediated gene expression profiles. We used MotifScanner to identify high scoring TFBS in the 1000bp upstream regions of the respective AR responsive genes and then identified overrepresented motif sites. As an alternative approach we applied a sequence-based $k$-mer feature search for motif detection. The distributions of all $k$-mers and known TFBS of DHT-regulated genes were used to characterize whether they are bound or not by AR. Overall, the best classifier achieves a sensitivity of 69% and a predictive accuracy of 69%.

After comparing the 5-mers with exactly one mismatch with known TFBS motifs detected in the AR-responsive gene group, the application of promoter classification linked known TFs into a model of AR-mediated gene expression and revealed transcriptional control mechanisms. The C/EBP and PAX motifs were identified as the strongest classifiers of androgen-regulated gene expression. Both of them were found in all 745 androgen-regulated genes and were significantly enriched in the promoters of genes encoding protein binding, ion binding and transferase activity. The PAX motif in the Pax4 promoter was previously shown to be involved in embryonic pattern formation, cell proliferation and cell differentiation (Sosa-Pineda 2004). The C/EBP binding
protein is a transcription factor that belongs to the basic leucine zipper protein family (Chattopadhyay 2006). It plays a crucial role in regulating balance between cell proliferation and differentiation through its recruitment onto ARE promoters, and is further able to inhibit expression of androgen-regulated genes, such as prostate-specific antigen.

The CREB motif is enriched in genes involved in various cellular processes including cell growth, survival and apoptosis. It has been demonstrated that interactions between AR and CREB through coactivator proteins including CBP/p300 regulate prostate specific antigen gene expression (Aarnisalo, 1998). A number of growth factors, hormones, and protein kinases including protein kinase A (PKA) and mitogen activated protein kinases (MAPKs) are involved in activating CREB. More importantly, it has been shown that interactions between CREB and AR signaling pathways may be one of the contributing factors involved in the development of androgen-independent prostate cancer (Garcia, Nicole et al. 2006). The GATA sites, which flank the AREs in the far upstream region of androgen responsive genes (Schuur et al., 1996; Cleutjens et al., 1997a; Pang et al., 1997; Zhang et al., 1997), are required for optimal gene expression and stimulated by androgen. Recent reports describe that PSA is a potential target gene for GATA transcription factors by regulating the PSA enhancer (Perez-Stable 2000). The interaction between GATA factor and AR are important for proper gene regulation in the prostate. In addition, GATA and AR-specific cofactors (Yeh and Chang, 1996; Tsang et al., 1997) present in different cell types may play a role in tissue-specific expression. GATA's principal role as a nuclear regulator may be to mark critical sites on DNA and to serve as a docking station for the assembly of a cell-specific transcriptional complex (Tsang 1997).

M00056-V$MYOGNF1_01 is the TFBS for myogenin or nuclear factor 1 transcription factor. Myogenic regulatory factors, including MyoD, myogenin, and other cell cycle regulatory factor genes are transcriptional targets of androgen regulation. Transient transfection studies
demonstrate that the role of the androgen signaling pathway is in regulation of myoblast cell growth and myogenic regulatory factors via enhanced myogenin expression (Lee 2002). E-box related factors ($DELTAEF$) which are enriched by helix-loop-helix transcription factor (bHLH) are involved in cell proliferation, differentiation and apoptosis. A recent study demonstrates that Pod-1, a member of the bHLH family, represses the activity of the AR promoter, as well as AR transactivation through its direct interaction with AR. Meanwhile, Pod-1 binds to the E-box element in the ARE promoter to repress gene transcription (Hong 2005). As forkhead transcription factors (M00269-V$XFD3_01$) widely participate in the development of various organs, it is reasonable to expect that they may play an important role in androgen signaling. Previous biochemical analysis identifies that forkhead transcription factor Foxa1 interacts with androgen signaling and controls prostate differentiated responses. Foxa1 plays a pivotal role in controlling prostate morphogenesis and cell differentiation (Gao 2005).

Integrating microarray expression profile and genome sequence data to identify co-regulated genes, followed by construction of transcriptional networks is currently the focus of many research groups. Our analysis of promoter sequences uses RF classifiers to predict AR responsive genes based on promoter sequence features. Knowing the direct, functional targets of androgens may increase our understanding of the different mechanisms by which AR acts and aids the development of directed therapies toward different regulatory modules of the androgen response (Phuc Le 2005). An understanding of the genes regulated by the androgen signaling pathway may also lead to more targeted therapies, thereby preventing unwanted side effects. Further work is likely to yield much more comprehensive and accurate models of the androgen regulatory networks of human and other organisms.
3.6 Acknowledgements

We are particularly indebted to Dmitri Kazmin and Donald P. McDonnell for the microarray data used in this paper. This research was funded by the Long-Range Research Initiative of the American Chemistry Council.

3.7 References


Figure 3.1. Random forests classification process. To generate the random forest, we select for each tree a subset of training data. Next, for every node in these trees a random subset of the attributes is chosen, and the attribute achieving the best division is selected.
Figure 3.2. The scatter plot results after display options have been set and up-regulated and down-regulated genes selected. The vertical position of each gene represents its expression level in the treatment of DHT at different concentrations, and the horizontal position represents its control condition.
Figure 3.3. F-scores are used for feature selection for mismatched 5-mer dataset. The dashed line depicts the manually chosen cutoff value below which features are excluded from training.
Table 3.1. Top ranked mismatched 5-mer motifs based on their importance along with the corresponding F-scores.

<table>
<thead>
<tr>
<th>Recognition sequence</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>tagcg_</td>
<td>0.0229</td>
</tr>
<tr>
<td>gcgc_c</td>
<td>0.0221</td>
</tr>
<tr>
<td>c_ccgg</td>
<td>0.0210</td>
</tr>
<tr>
<td>cg_tcc</td>
<td>0.0198</td>
</tr>
<tr>
<td>cgctc_</td>
<td>0.0193</td>
</tr>
<tr>
<td>ggcg_t</td>
<td>0.0191</td>
</tr>
<tr>
<td>t_gcg</td>
<td>0.0189</td>
</tr>
<tr>
<td>cac_tg</td>
<td>0.0183</td>
</tr>
<tr>
<td>aaa_cg</td>
<td>0.0183</td>
</tr>
<tr>
<td>gc_ccg</td>
<td>0.0182</td>
</tr>
<tr>
<td>g_ctgt</td>
<td>0.0180</td>
</tr>
<tr>
<td>gtg_ct</td>
<td>0.0179</td>
</tr>
<tr>
<td>cgcc_cc</td>
<td>0.0177</td>
</tr>
<tr>
<td>tgac_g</td>
<td>0.0172</td>
</tr>
<tr>
<td>ctga_g</td>
<td>0.0172</td>
</tr>
<tr>
<td>gg_tgt</td>
<td>0.0171</td>
</tr>
<tr>
<td>cgccg_c</td>
<td>0.0170</td>
</tr>
<tr>
<td>gcgcc_c</td>
<td>0.0169</td>
</tr>
<tr>
<td>ct_tac</td>
<td>0.0166</td>
</tr>
<tr>
<td>c_gtga</td>
<td>0.0165</td>
</tr>
</tbody>
</table>
Figure 3.4. Screenshot of Toucan 2 with known TFBS analysis result. The entire set of human motifs in TRANSFAC 7.0 was used for MotifScanner. Each input gene is displayed as a straight line whose length is identical to the length of the original sequence in nucleotides. Individual TF binding sites are represented as colored rectangular blocks.
Figure 3.5. Curves of F-scores against TFBS features. F-score selection of top ranking TFBS in DHT-regulated genes.
Table 3.2. Top ranked TFBS names, recognition sequences, and their corresponding F-score values in order of importance.

<table>
<thead>
<tr>
<th>TFBSs ID</th>
<th>Sequences</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>M00377-V$PAX4_02</td>
<td>GAAAAAATAAC</td>
<td>0.5194</td>
</tr>
<tr>
<td>M00073-V$DELTAEF1_01</td>
<td>TTTCACCTTGT</td>
<td>0.4380</td>
</tr>
<tr>
<td>M00269-V$XFD3_01</td>
<td>AGATTAAATAAACT</td>
<td>0.3967</td>
</tr>
<tr>
<td>M00078-V$EVI1_01</td>
<td>TGAGAAGATTAATAA</td>
<td>0.3479</td>
</tr>
<tr>
<td>M00100-V$CDXA_01</td>
<td>ATTTATA</td>
<td>0.3416</td>
</tr>
<tr>
<td>M00190-V$CEBP_Q2</td>
<td>ATCTTGAATAAATT</td>
<td>0.3276</td>
</tr>
<tr>
<td>M00114-V$TAXCREB_01</td>
<td>GGGGGATGACGTCAC</td>
<td>0.3030</td>
</tr>
<tr>
<td>M00126-V$GATA1_02</td>
<td>TCATTGATAATAGA</td>
<td>0.2957</td>
</tr>
<tr>
<td>M00056-V$MYOGNF1_01</td>
<td>AAGATTTTATGTTGGATTTTGCCCAGAA</td>
<td>0.2868</td>
</tr>
<tr>
<td>M00040-V$CREBP1_01</td>
<td>TTACATAA</td>
<td>0.2821</td>
</tr>
<tr>
<td>M00473-V$FOXO1_01</td>
<td>AATAAACAGG</td>
<td>0.2742</td>
</tr>
<tr>
<td>M00129-V$HFH1_01</td>
<td>TATTGTTTAGTT</td>
<td>0.2682</td>
</tr>
<tr>
<td>M00074-V$CETS1P54_02</td>
<td>ACAAAGGAAAAATG</td>
<td>0.2485</td>
</tr>
<tr>
<td>M00258-V$ISRE_01</td>
<td>CGGTTCGTTTCTCT</td>
<td>0.2431</td>
</tr>
<tr>
<td>M00462-V$GATA6_01</td>
<td>AAAGATAAAA</td>
<td>0.2297</td>
</tr>
<tr>
<td>M00001-V$MYOD_01</td>
<td>GGTCAGCTGTCG</td>
<td>0.2125</td>
</tr>
<tr>
<td>M00189-V$AP2_Q6</td>
<td>CTCCCCCAGGCC</td>
<td>0.2071</td>
</tr>
<tr>
<td>M00396-V$EN1_01</td>
<td>ATATTTG</td>
<td>0.1820</td>
</tr>
<tr>
<td>M00007-V$ELK1_01</td>
<td>CAAAAAGGAAAGTAACT</td>
<td>0.1812</td>
</tr>
<tr>
<td>M00113-V$CREB_02</td>
<td>GGGATGACGTC</td>
<td>0.1767</td>
</tr>
</tbody>
</table>
Table 3.3. The prediction results using different k-mers: 4-mers, 5-mers, mismatched 4-mers and mismatched 5-mers.

<table>
<thead>
<tr>
<th></th>
<th>4-mer without gap motif</th>
<th>4-mer with gap motif</th>
<th>5-mer without gap motif</th>
<th>5-mer with gap motif</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OOB error rate</strong></td>
<td>40.2%</td>
<td>39.6%</td>
<td>39.7%</td>
<td>36.11%</td>
</tr>
<tr>
<td></td>
<td>39.2%</td>
<td>39.33%</td>
<td>39.3%</td>
<td>36.24%</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.6</td>
<td>0.61</td>
<td>0.6</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.61</td>
<td>0.61</td>
<td>0.65</td>
</tr>
<tr>
<td><strong>True Positive</strong></td>
<td>0.61 (451/745)</td>
<td>0.65 (481/745)</td>
<td>0.64 (478/745)</td>
<td>0.67 (492/745)</td>
</tr>
<tr>
<td><strong>Prediction Rate</strong></td>
<td>0.65 (483/745)</td>
<td>0.65 (483/745)</td>
<td>0.65 (483/745)</td>
<td>0.68 (506/745)</td>
</tr>
<tr>
<td><strong>True Negative</strong></td>
<td>0.6 (440/745)</td>
<td>0.57 (419/745)</td>
<td>0.57 (421/745)</td>
<td>0.62 (460/745)</td>
</tr>
<tr>
<td><strong>Prediction Rate</strong></td>
<td>0.57 (422/745)</td>
<td>0.57 (421/745)</td>
<td>0.57 (421/745)</td>
<td>0.61 (454/745)</td>
</tr>
</tbody>
</table>
Table 3.4. Performance comparison of the RF classifier of TFBS with a combination of TFBS and mismatched 5-mer features in DHT-regulated gene expression.

<table>
<thead>
<tr>
<th></th>
<th>TFBS motifs</th>
<th>TFBS motifs and 5mer motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OOB error rate</strong></td>
<td>36.44%</td>
<td>33.83%</td>
</tr>
<tr>
<td></td>
<td>36.64%</td>
<td>33.97%</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.636</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>0.634</td>
<td>0.657</td>
</tr>
<tr>
<td><strong>True Positive Prediction Rate</strong></td>
<td>0.64 (476/745)</td>
<td>0.70 (518/745)</td>
</tr>
<tr>
<td></td>
<td>0.635 (473/745)</td>
<td>0.70 (519/745)</td>
</tr>
<tr>
<td><strong>True Negative Prediction Rate</strong></td>
<td>0.632 (471/745)</td>
<td>0.68 (506/745)</td>
</tr>
<tr>
<td></td>
<td>0.632 (471/745)</td>
<td>0.62 (460/745)</td>
</tr>
</tbody>
</table>
Table 3.5. Comparison of mismatched 5-mer features and the TFBS motifs. Highlights in red are the perfect matches between top-ranked 5-mers and top-ranked TFBS motifs.

<table>
<thead>
<tr>
<th>TFBSs ID</th>
<th>Sequences</th>
<th>Motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>M00377-V$PAX4_02</td>
<td>GAAAAAATAAC CAAAAATTAGC</td>
<td>PAX</td>
</tr>
<tr>
<td>M00073-V$DELAEF1_01</td>
<td>CTTCACCTGGT TTTCACCTTGT</td>
<td>EBOX</td>
</tr>
<tr>
<td>M00269-V$XF3D_01</td>
<td>TCTGTGAAATAATA TATGTAATAAAACT</td>
<td>Fork Head Domain factors</td>
</tr>
<tr>
<td>M00078-V$EVI1_01</td>
<td>TGAGAAGATTAATATA GGAGAAGCCTAGATAA</td>
<td>yeloid transforming protein</td>
</tr>
<tr>
<td>M00100-V$CDX_01</td>
<td>CTTTATA ATTTATA</td>
<td>caudal related homeodomain protein</td>
</tr>
<tr>
<td>M00190-V$CEBP_Q2</td>
<td>ATCTTGAAATAATT</td>
<td>C/EBP</td>
</tr>
<tr>
<td>M00114-V$TAXCREB_01</td>
<td>TGGCGCTGACGACGA AATGGCTGACGTCTA</td>
<td>CREB</td>
</tr>
<tr>
<td>M00126-V$GATA1_02</td>
<td>TCATTGATAATAGA</td>
<td>GATA</td>
</tr>
<tr>
<td>M00056-V$MYO1N1_01</td>
<td>CACTGCTTTAGTCTGGCCTCTGCCCTGCC TGACTGTCTTTTGATTCGCAAAACTCAGCA</td>
<td>MYOGENIN</td>
</tr>
<tr>
<td>M00040-V$CREBP1_01</td>
<td>TGACGTTA TGACATAA</td>
<td>CREB</td>
</tr>
</tbody>
</table>
Chapter 4

Non-monotonic Dose-response Relationship in Steroid Hormone Receptor-mediated Gene Expression

Li Li, Melvin E. Andersen, Steffen Heber, and Qiang Zhang
4.1 Abstract
Steroid hormone receptors are the targets of many environmental endocrine active chemicals (EACs) and synthetic drugs used in hormone therapy. While most of these chemical compounds have a unidirectional and monotonic effect, certain EACs can display non-monotonic dose–response behaviors and some synthetic drugs are selective endocrine modulators. Mechanisms underlying these complex endocrine behaviors have not been fully understood. By formulating an ordinary differential equation-based computational model, we investigated in this study the steady-state dose–response behavior of exogenous steroid ligands in an endogenous hormonal background under various parameter conditions. Our simulation revealed that non-monotonic dose–responses in gene expression can arise within the classical genomic framework of steroid signaling. Specifically, when the exogenous ligand is an agonist, a U-shaped dose–response appears as a result of the inherently nonlinear process of receptor homodimerization. This U-shaped dose–response curve can be further modulated by mixed-ligand heterodimers formed between endogenous ligand bound and exogenous ligand-bound receptor monomers. When the heterodimer is transcriptionally inactive or repressive, the magnitude of U-shape increases; conversely, when the heterodimer is transcriptionally active, the magnitude of U-shape decreases. Additionally, we found that an inverted U-shaped dose–response can arise when the heterodimer is a strong transcription activator regardless of whether the exogenous ligand is an agonist or antagonist. Our work provides a novel mechanism for non-monotonic, particularly U-shaped, dose–response behaviors observed with certain steroid mimics, and may help not only understand how selective steroid receptor modulators work but also improve risk assessment for EACs.

Key words: steroid hormone receptor, non-monotonic, U-shape, homodimerization, heterodimer.
4.2 Introduction

Steroid hormone receptors (SHRs) comprise a superfamily of transcription factors that are activated by steroid hormones to regulate specific gene expression. They play critical roles in a variety of physiological processes including homeostasis, metabolism, reproduction, and behaviors. Widely studied SHRs include androgen receptor (AR), estrogen receptor (ER), progesterone receptor (PR), and glucocorticoid receptor (GR). The overall structure of these SHRs is highly conserved and contains four major functional domains: a conserved DNA-binding domain (DBD), a C-terminal ligand-binding domain (LBD), an N-terminal transactivational domain (NTD), and a hinge region connecting DBD and LBD (Ruff, Gangloff et al. 2000; Gelmann 2002; Bledsoe, Stewart et al. 2004; Rogerson, Brennan et al. 2004). In its most general form, the classical genomic action of steroid hormones involves the following intracellular processes. In the absence of natural ligand, SHRs in the cytoplasm or nucleus are associated with chaperone complexes, which function to stabilize SHRs and enhance their ligand binding affinity (Fang, Fliss et al. 1996; Pratt and Toft 1997). During ligand binding, SHRs first dissociate from the chaperone complexes, followed by nuclear translocation of those SHRs initially located in the cytoplasm (Htun, Barsony et al. 1996; Georget, Lobaccaro et al. 1997; Tyagi, Lavrovsky et al. 2000). Liganded receptor monomers then homodimerize with each other to form receptor dimers (Kumar and Chambon 1988; Wrangle, Eriksson et al. 1989; Wang, Peters et al. 1995). Depending on the type of SHRs, receptor dimers recognize and bind specifically to the DNA of target hormone response elements (HREs), with each monomer recognizing one half-site of the HRE (Nordeen, Suh et al. 1990; Langley, Zhou et al. 1995; Klinge, Silver et al. 1997). The receptor dimer-DNA complex then recruits a battery of nuclear coregulators to alter the local chromatin structures (Heinlein and Chang 2002; Smith and O'Malley 2004). With a relaxed chromatin structure, polymerase II gains access to the promoter to initiate gene transcription. Ligands that allow recruitment of corepressors can induce condensation of the chromatin, thus turning off gene transcription (Fernandes and White 2003).
SHRs that are normally activated by endogenous ligands can also respond to exogenous substances, including synthetic drugs and environmental chemicals. Acting on an endogenous hormonal background, these chemical compounds can potentially modulate endocrine events, resulting in altered SHR-mediated biological functions. Many therapeutic drugs are designed to interact directly with SHRs as agonists or antagonists to alter gene transcriptional activities in target tissues. The effects exerted by many of these drugs may vary from tissue to tissue (Dutertre and Smith 2000; Giannoukos, Szapary et al. 2001; Berrevoets, Umar et al. 2002). For example, while both tamoxifen and raloxifene both reduce the risk of invasive breast cancers by acting as ER antagonists in mammary tissues, they have a strong agonist activity in bones that helps maintain bone density in women (Dutertre and Smith 2000; Francucci, Romagni et al. 2005). Because of the opposing effects, these drugs are more appropriately termed as selective receptor modulators (SRMs). Besides synthetic drugs, a large set of chemicals that can interfere with endocrine functions are environmental pollutants termed as endocrine active chemicals (EACs). EACs may interfere with the synthesis and metabolism of endogenous hormones, or in many cases, interact with SHRs directly (Amaral Mendes 2002; Markey, Rubin et al. 2002). An important aspect of health risk assessment for EACs is understanding dose-response curves at low doses that are relevant to human exposure in the environment. A large body of evidence indicates that SHR-mediated adverse effects of EACs are sometimes nonlinear or even non-monotonic (i.e., U-shaped or inverted U-shaped) in dose ranges exerting no overt cytotoxicity (Kemppainen and Wilson 1996; vom Saal, Timms et al. 1997; Maness, McDonnell et al. 1998; Putz, Schwartz et al. 2001; Putz, Schwartz et al. 2001; Almstrup, Fernandez et al. 2002; Terouanne, Nirde et al. 2003; Kohlerova and Skarda 2004).

The molecular basis for the bidirectional actions of SRMs and non-monotonic or hormetic effects of EACs are not completely understood. A variety of mechanisms have been proposed to explain
these observations. For instance, the ratio of coactivators (CoA) and corepressors in a cell may determine whether an exogenous ligand behaves primarily as an agonist or antagonist (Smith, Nawaz et al. 1997; Szapary, Huang et al. 1999; Smith and O'Malley 2004). Alternatively, the opposing effect of SRMs in different tissues may result from involvement of different SHR subtypes (Mclnerney, Weis et al. 1998; Zhou and Cidlowski 2005). Kohn and Melnick found that an inverted U-shaped dose-response can arise from conditions where there are unoccupied receptors by endogenous hormones and recruitment of CoAs by xenobiotic ligands is relatively weak (Kohn and Melnick 2002). Conolly and Lutz hypothesized that a U-shaped dose-response curve can result from transcriptional inactive mixed-ligand receptor dimers (Conolly and Lutz 2004). Possibilities also exist that the non-genomic effect of steroid ligands, which often leads to activation of kinases such as mitogen-activated protein kinase, may modulate the genomic actions of the same ligands in opposite directions via receptor or coregulator phosphorylation, resulting in non-monotonic responses in gene expression (Acconcia and Marino 2003; Rochette-Egly 2003).

The present study focused on the steady-state dose-response for gene expression mediated through SHRs. Using a computational modeling approach we demonstrated that non-monotonic dose-responses can readily arise within the classical framework of steroid signaling. Our results indicated that the inherently nonlinear process of receptor homodimerization in SHR signaling plays an important role in rendering U-shaped dose-response curves, which can be further modulated by mixed-ligand heterodimers.

4.3 Methods

4.3.1 Model Structure

Definitions for pure agonist, antagonist, and partial agonist in particular, in the endocrine literature have been largely observational rather than mechanistic. For modeling purposes we need to be
more explicit, and so these terms are defined as follows. A pure agonist is a ligand that is able to recruit CoA exclusively to activate gene transcription. If a ligand is able to recruit corepressors (CoR) exclusively to deactivate gene transcription, it is termed as an active antagonist, whereas if a ligand can recruit neither CoR nor CoA after binding to a steroid receptor, it is termed as a passive antagonist. A partial agonist is a ligand that is able to recruit both CoA and CoR, albeit not simultaneously. In this way, when a partial agonist acts alone in our model, it always activates gene transcription but with a reduced maximal response compared with a pure agonist. It is generally believed that whether a receptor is able to recruit CoA or CoR depends on its conformational changes after ligand binding, particularly in the LBD domain (Brzozowski, Pike et al. 1997; Shiau, Barstad et al. 1998), and the phosphorylation status of the receptor also modulates this recruiting process (Atanaskova, Keshamouni et al. 2002; Rochette-Egly 2003).

Since there are always background levels of endogenous hormones in a physiological state, we simulated gene expression driven by exogenous ligand X in the presence of endogenous ligand L (Figure 4.1). In the absence of X, endogenous ligand L first binds to receptor SHR to form a liganded receptor complex LR. Two LRs then associate with each other to form a receptor homodimer LRRL. LRRL in turn binds to the HRE in the promoter of target genes. While bound to HRE, LRRL is able to recruit CoA to the local promoter site and together these molecules produce an activational complex CoALRRLH. Since L mimics an endogenous hormone here, we assumed that L acts only as a pure agonist, thus by our definition, recruiting no CoR to the local promoter. Exogenous ligand X follows a similar signaling process as the endogenous ligand L. However, X may function as a pure agonist by recruiting CoA, a partial agonist by recruiting both CoA and CoR, or a passive or active antagonist. When both endogenous ligand L and exogenous ligand X are present, it is also possible that L-bound receptor LR may interact with X-bound receptor XR to form mixed-ligand heterodimers (LRRX). Similar to XRRX, LRRX may regulate gene transcription differentially, depending on whether CoA, CoR, or both are recruited. Although
Figure 4.1 indicates that recruitment of CoA or CoR occurs after a receptor dimer binds to HRE, the dimer may also interact with CoA or CoR directly prior to occupying HRE (Thenot, Bonnet et al. 1999; Margeat, Poujol et al. 2001). We found that inclusion of these DNA-independent interactions between receptor dimers and coregulators did not qualitatively change the simulation results obtained in the absence of these interactions, except for the circumstance of receptor overexpression, in which excessive receptors may serve to scavenge the free coregulators resulting in repression of gene expression at high doses of X. This auto-inhibitory effect of receptor overexpression has been observed in vitro with ERs (Bocquel, Kumar et al. 1989; Webb, Lopez et al. 1992). Therefore the present study only presents results considering DNA-dependent recruitment.

The control of gene activation at the promoter was modeled based on the current understanding of gene induction in eukaryotic cells (Zhang, Andersen et al. 2006). At any given time, a gene could be in one of two discrete transcriptional states, inactive (GENE$_i$) or active (GENE$_a$), corresponding to compact and relaxed chromatin structures, respectively. Once in the active state, gene transcription proceeds at a relatively constant rate, whereas in the inactive state, no transcription occurs. Transitions between the inactive and active states are controlled by rate constant $k_{f5}$ and $k_{b5}$ as indicated below,

$$\text{GENE}_i \xrightarrow{k_{f5}} \text{GENE}_a \xleftarrow{k_{b5}} \text{GENE}_i,$$

where

$$k_{f5} = k_{f51}[\text{CoALRRLH}] + k_{f52}[\text{CoAXRXHX}] + k_{f53}[\text{CoALRXHX}], \quad (1)$$

$$k_{b5} = k_{b51} + k_{b52}[\text{CoRXRXHX}] + k_{b53}[\text{CoRLRXHX}]. \quad (2)$$

Notably, the transition from the inactive to active state is regulated by coactivator-bound receptor-DNA complexes CoALRRLH, CoAXRXHX, and CoALRXHX (Eq. (1)). These complexes would work to relax local chromatin structures by acting as or recruiting acetyltransferase, a process not
explicitly modeled. Conversely, transition from the active to inactive state is regulated by corepressor-bound receptor-DNA complexes CoRXRRXH and CoRLRRXH (Eq. (2)), which presumably convert relaxed chromatin into a compact structure by acting as or recruiting histone deacetylase. The term $K_{b51}$ serves as a constitutive repressor activity, which turns off gene transcription in the absence of ligand-induced corepressor recruitment. Once in the active state, the gene transcribes primary transcripts (PTs). PTs are processed to become mature mRNAs, followed by protein translation. By modeling the process of gene regulation with these steps, we were able to incorporate both positive and negative transcriptional controls in a mechanistically more accurate manner, rather than relying on empirical equations.

### 4.3.2 Model Parameters

Ordinary differential equations (ODEs) and parameter values are listed in the Supporting Material (Table 4.S1 and 4.S2), including references and rationale for the choice of parameter values. For direct comparison, the default parameter values for exogenous ligand X-initiated processes and mixed-ligand heterodimer-initiated processes were set the same as for endogenous ligand L. But these parameters were varied systemically in the preset study to investigate their effects on dose-response curves. Since we are interested only in the steady-state dose-response behavior, only the forward association rate constants of reversible reactions were varied to investigate the effects of these processes.

### 4.3.3 Modeling Tools

The computational model was first constructed and parameterized in PathwayLab (InNetics, Inc., Sweden), then exported into MatLab (The Mathworks, Inc., Natick, MA). Dose-response curves were obtained by running the model to steady state in MatLab. The model in the System Biology Markup Language (SBML) and MatLab format is provided as additional supplementary materials.
4.4 Results
Since exogenous ligand X can be either a pure agonist, an antagonist, or a partial agonist, we explored the steady-state dose-response relationship between gene expression and X for each of the three possibilities in this order. Moreover, with each possibility, we also considered situations in which mixed-ligand heterodimers either do not form at all, act as an activator, or as a repressor.

4.4.1 Exogenous Ligand X as a Pure Agonist
4.4.1.1 In the Absence of Mixed-ligand Heterodimer LRRX
Contrary to the intuition that an agonistic exogenous ligand X would add to the basal gene expression sustained by endogenous ligand L, simulations surprisingly revealed that X, acting on top of L, exhibits non-monotonic U-shaped dose-response curves (Figure 4.2, left panels). X at relatively low doses first depresses the basal gene expression, and after reaching a minimum expression, the steady-state protein level reverses the downtrend as the dose of X continues to increase and finally reaches a saturated phase. This U-shaped profile was preserved in most of the conditions where parameter values associated with the signaling events were varied (see details below). To quantify the U-shape, we regard the magnitude of U-shape as the difference between the expression level at the nadir and the lesser of the basal and saturated expression levels. Conversely, the magnitude of inverted U-shape is the difference between the expression level at the peak and the greater of the basal and saturated expression levels. For continuity, the difference in positive values denotes a U-shaped response, negative values an inverted U-shaped response.

Since the physiological level of an endogenous steroid hormone (represented by L), such as testosterone and estrogen, varies between individuals or fluctuates through various physiological states, we investigated the dose-response curve for exogenous ligand X in the presence of
different L levels. As indicated in Figure 4.2A, with higher L levels, the nadir of the U-shaped
dose-response curve shifts progressively to the right, indicating increasing difficulty for X to
initially repress gene expression activated by L. However, with higher L levels, the magnitude of
U-shape becomes more prominent, although it eventually levels off (Figure 4.2A, right panel).
Exogenous ligands often differ in their binding affinity towards target SHRs. By varying $k_{f02}$, the
association rate constant between X and SHR, our simulation demonstrated that an increase in
binding affinity merely results in a parallel, leftward shift of the dose-response curve without
affecting the magnitude of U-shape (Figure 4.2B). Another important variable is the intracellular
concentration of SHRs, which may be at different levels among different cell types and tissues
(kuiper et al. 1997) or at different developmental and physiological stages, thus affecting cellular
responses to endogenous hormones and exogenous ligands. By increasing the initial abundance
of SHR in the model, we found that the dose-response curve generally shifts upward, with the
dose of X associated with the expression nadir remaining largely unchanged (Figure 4.2C).
Increasing the abundance of SHR initially enhances the magnitude of U-shape, which then
diminishes as SHR increases further. Next, we examined how the ability of receptor monomer XR
to form homodimer XRRX affects the dose-response curve. With a low association rate constant
($k_{r12}$) for homodimerization, exogenous ligand X behaves almost as a pure antagonist (Figure
4.2D). In contrast, with a high $k_{r12}$ value, X produces a complete agonistic effect. With
intermediate $k_{r12}$ values, however, the dose-response curve remains U-shaped. Similar to the
effect of varying SHR abundance, the magnitude of U-shape also has a biphasic appearance
(Figure 4.2D, right panel). Lastly, we investigated how the ability of receptor dimer XRRX to bind
HRE ($k_{f22}$), XRRXH to recruit CoA ($k_{f32}$), and CoAXRRXH to activate $\text{GENE}_r$-$\text{GENE}_a$ transition
($k_{f52}$), affects the shape of the dose-response curve. Variation in these parameters gives rise to a
similar curvature change as varying $k_{r12}$ (results not shown).
With the above analyses, it appears that the U-shaped profile of dose-response persists in most of the situations explored. Varying parameter values at different stages of the signaling pathway seems to affect, in most cases, only the magnitude and/or position of the U-shape, rather than completely eradicate it. To identify the origin of the U-shaped dose-response, we then focused on the step of homodimerization between receptor monomers. This is an inherently nonlinear process, with a quadratic term describing the forward association rate, as indicated in Equation (3) and (4),

\[ f_{flux11} = k_{f11}[LR]^2, \quad (3) \]
\[ f_{flux12} = k_{f12}[XR]^2. \quad (4) \]

Linearizing the dimerization processes by converting Equation (3) and (4) into (5) and (6), respectively,

\[ f_{flux11} = k'_{f11}[LR], \quad (5) \]
\[ f_{flux12} = k'_{f12}[XR]. \quad (6) \]

(where \( k'_{f11} \) and \( k'_{f12} \) were set to maintain the same basal gene expression level), we found that the U-shape was completely eliminated, and under no circumstances did it recur by varying parameter values in any of the signaling steps (Figure 4.3). These results indicated that the U-shaped response must originate from receptor homodimerization, and it may be understood as follows. When X is competing against L for SHR at a low dose, the loss of homodimer LRRL from this competition cannot be fully compensated by newly formed XRRX due to the nonlinearity inherent in homodimerization. This inability to replenish lost LRRL with XRRX results in an initial depression in gene expression. At a higher dose of X, more XRRX will be formed, which is eventually high enough to compensate for all the losses of LRRL, thus reversing the downtrend in gene expression.
4.4.1.2 In the Presence of Mixed-ligand Heterodimer LRRX

In this case we considered situations in which LRRX acts as either a transcriptional activator by recruiting CoA, an active repressor by recruiting CoR, a passive repressor by not binding to HRE or recruiting any coregulators, or a partial activator by recruiting both CoA and CoR, though not simultaneously.

4.4.1.2.1 LRRX as an Activator

Compared with the situation devoid of heterodimer formation (i.e., $k_{13}=0$), emergence of LRRX as a transcriptional activator attenuates the magnitude of U-shape. As $k_{13}$, the association rate constant between LR and XR, increases, more LRRX heterodimers are formed to activate gene expression, pushing the nadir of the U-shaped dose-response curve upward and thereby reducing the magnitude of U-shape (Figure 4.4A, top panel). When $k_{13}$ reaches a value comparable to the equivalent parameters in the homodimerization processes (i.e., $k_{11}$ and $k_{12}$), the U-shape essentially disappears and the response only increases monotonically. The dose-response curve remains monotonic within a range of $k_{13}$, as indicated by the extended horizontal line at zero in Figure 4.4A (middle panel). As $k_{13}$ increases further, the dose-response curve leaves the monotonic bounds and appears non-monotonic again. Instead of a U-shape, an inverted U-shape emerges in this case, and its magnitude, as represented by negative values, increases sharply with small increments of $k_{13}$. Notably, the overall shape of the dose-response curve is the sum of contributions from both homodimers LRLR and XRXR, and the heterodimer LRRX (Figure 4.4A bottom panel). At a high $k_{13}$ value, an inverted U-shaped dose-response curve results because LRRX, which itself is inverted U-shaped in appearance, has a dominant influence. Varying the association rate constant for LRRXH to recruit CoA ($k_{33}$) has a modulatory effect similar to $k_{13}$ on the steady-state dose-response (Figure 4.4B). However, when $k_{33}$ is too low, the magnitude of U-shape has a tendency to increase as LRRX essentially degenerates to a passive repressor. Variations in the association rate constant for LRRX to bind HRE ($k_{23}$) and for
CoALRRXH to activate GENE$_{i}$-to-GENE$_{a}$ transition ($k_{53}$) have an effect similar to varying $k_{33}$ (results not shown). Overall, the inverted U-shaped curve originates from the formation of LRRX, which play a dominant role in gene expression when they are in high abundance or are highly transcriptionally active.

4.4.1.2.2 LRRX as a Repressor

If the mixed-ligand heterodimer LRRX does not recruit CoA, it functions as a transcriptional repressor, either active or passive, depending on whether CoR can be recruited. If LRRX does not bind to the response element HRE, or after binding, does not recruit CoR, then LRRX would behave as a passive repressor by making LR and XR less available for formation of homodimers, or by making HRE less available to homodimers which activate gene expression. If LRRX recruits CoR after binding to HRE, then LRRX acts as an active repressor by promoting transition from GENE$_{a}$ to GENE$_{i}$. Simulations revealed that regardless of being passive or active, existence of LRRX as a repressor further deepens the U-shaped dose-response curve observed in the absence of LRRX formation (Figure 4.5). When LRRX is mimicked as a passive repressor, which cannot bind to HRE (i.e., $k_{23}=0$), increasing $k_{13}$, the association rate constant between LR and XR, progressively enhances the magnitude of U-shape, with the nadir of the U-shape eventually dropping to zero expression level (Figure 4.5A). Similarly, mimicking LRRX as a passive repressor that can bind to HRE but unable to recruit CoR (i.e., $k_{43}=0$) also produces a U-shape deepening effect (results not shown). When LRRX acts as an active repressor, increasing $k_{43}$ (which represents the ability of LRRXH to recruit CoR) augments the magnitude of U-shape, as well (Figure 4.5B).

4.4.1.2.3 LRRX as a Partial Activator

When LRRX is able to recruit both CoA and CoR, thus acting as a partial activator by itself, its effect on regulating the direction of gene expression in the presence of endogenous ligand L
depends on the relative strength of activation and repression. If LRRX recruits CoA more strongly than CoR, the shape of dose-response curves will range from blunted U-shape to inverted U-shape (results not shown), similar to the effect of LRRX acting as a pure activator. Conversely, if LRRX recruits CoR more strongly than CoA, a deepened U-shaped dose-response curve arises (results not shown), similar to the effect of LRRX acting as a repressor.

4.4.2 Exogenous Ligand X as an Antagonist

4.4.2.1 In the Absence of Mixed-ligand Heterodimer LRRX

As an antagonist, X may repress gene expression either passively or actively. With passive repression, X competes against L for receptors or response elements; with active repression, X recruits CoR to promote deactivation of actively transcribing genes. Simulations revealed that as an antagonist, regardless of being passive or active, X produces monotonically decreasing responses in gene expression (Figure 4.6). Increasing the binding affinity between X and SHR by adjusting \(k_{f02}\) shifts the dose-response curve to the left in parallel (Figure 4.6A). In comparison, increasing the association rate constants in downstream steps (i.e., \(k_{f12}, k_{f22}, k_{f42}\)) not only shifts the dose-response curve to the left, but also steepens the monotonically decreasing slope (Figure 4.6B-D). In no circumstances was a non-monotonic dose-response curve, either U- or inverted U-shaped, observed.

4.4.2.2 In the Presence of Mixed-ligand Heterodimer LRRX

4.4.2.2.1 LRRX as an Activator

By acting as an activator, LRRX alleviates the antagonistic action of X (Figure 4.7). Increasing \(k_{f13}\), the association rate constant between LR and XR, initially shifts the monotonically decreasing dose-response curve to the right without changing its monotonic nature (Figure 4.7A). As \(k_{f13}\) increases further, the activity of LRRX starts to dominate the shape of the dose-response curve, resulting in an inverted U-shape, the magnitude of which (in negative values) increases
sharply for small increment of $k_{f13}$. Similar to the effect of varying $k_{f13}$, emergence of the inverted U-shape was also obtained by varying the following constants: $k_{f23}$, the association rate constant between LRRX and HRE (results not shown), by varying $k_{f33}$, the association rate constant for LRRXH to recruit CoA (Figure 4.7B), and $k_{f53}$, the constant for CoALRRXH to activate $GENE_i$-to-$GENE_a$ transition (results not shown). In all cases, high doses of X eventually repress the gene expression completely after an expression peak. This behavior is in contrast to the situation in which X is a pure agonist and it causes maximal expression at high doses instead of depressing it to zero (Figure 4.7 vs. Figure 4.4).

4.4.2.2.2 LRRX as a Repressor

By acting as a repressor, either passive or active, LRRX further strengthens the antagonistic action of X (Figure 4.8). Increasing $k_{f13}$, $k_{f23}$ (results not shown), $k_{f43}$, and $k_{f53}$ (results not shown), which enhances LRRX as a repressor at different signaling stages, results in leftward shifting of the monotonically decreasing dose-response curve. But unlike the case where X functions as an antagonist in the absence of heterodimer LRRX, the slope of the dose-response curves tends to decrease in steepness as they shift to the left. (Figure 4.8 versus. Figure 4.6).

4.4.2.2.3 LRRX as a Partial Activator

When LRRX is able to recruit both CoA and CoR, thus acting as a partial activator by itself, its effect on regulating the direction of gene expression depends on the relative strength of activation and repression. If LRRX recruits CoA more strongly than CoR, the shape of dose-response curves will range from monotonic decreasing to inverted U-shaped (results not shown), similar to the effect of LRRX acting as a pure activator. Conversely, if LRRX recruits CoR more strongly than CoA, only monotonically decreasing dose-response curve can be observed (results not shown), similar to the effect of LRRX acting as a repressor.
4.4.3. Exogenous Ligand X as a Partial Agonist

If exogenous ligand X is a partial agonist, then by our definition both CoA and CoR can be recruited, albeit not simultaneously, by XRRX occupying the response element HRE. A series of simulations revealed that when heterodimer LRRX is absent or acts as a repressor, U-shaped dose-response curves can arise (Figure 4.9A and B). When LRRX acts as a pure or partial activator, both U-shaped and inverted U-shaped responses can be observed, depending on its abundance and strength of activity (Figure 4.9C and D). The ratio between intracellular CoR and CoA has been proposed to explain the differential effects in different target tissues of many exogenous steroid mimics acting through SHRs (Smith, Nawaz et al. 1997; Smith and O'Malley 2004). Our simulation indicated that the CoR/CoA ratio does indeed affect the dose-response behavior in a very sensitive manner (Figure 4.9). An increase in CoR/CoA ratio tends to render the action of X completely antagonistic, whereas a decrease in the ratio makes X behave more like an agonist. Therefore, the relative abundance of CoR and CoA is a key modulator of SHR-mediated gene expression,

In summary, we have simulated the steady-state gene expression in an endogenous hormonal background with varying assumptions about the exogenous steroid ligand. The shape of dose-response curves can vary from monotonically increasing or decreasing to non-monotonically U-shaped or inverted U-shaped, depending on the transcriptional nature of the exogenous ligand and the intermediate complexes formed with the endogenous ligand. Conditions under which non-monotonic dose-responses could occur are summarized in Table 4.1. U-shaped dose-response curves may arise when exogenous ligand X is either a pure or partial agonist, regardless of the presence of mixed-ligand heterodimer LRRX; while inverted U-shape arises only when LRRX functions as a pure or partial activator, regardless of whether X by itself is an agonist or antagonist.
4.5 Discussion

The biological responses invoked by an exogenous chemical may be non-monotonic (Calabrese and Baldwin 2001; Calabrese and Baldwin 2003). With respect to SHR-mediated action, a variety of mechanisms have been proposed to explain the non-monotonic effects of certain EACs and the bidirectional effects of SHRs in different target tissues (Smith, Nawaz et al. 1997; McInerney, Weis et al. 1998; Kohn and Melnick 2002; Conolly and Lutz 2004; Smith and O'Malley 2004). Using numerical simulation of a relatively standard model of steroid hormone action, the present study demonstrated that non-monotonic steady-state response may be a property intrinsic to SHR-mediated gene expression under a variety of conditions.

An essential step in the genomic action of steroid hormones is the dimerization of liganded receptor monomers (Kumar and Chambon 1988; Wrangle, Eriksson et al. 1989). The requirement for SHRs to function as a dimer lies at least in the fact that a steroid HRE invariably comprises two half-sites of either direct or inverted repeats, and each monomer can only recognize one half-site weakly (Nordeen, Suh et al. 1990; Langley, Zhou et al. 1995; Kuntz and Shapiro 1997). Therefore, SHRs need to function as a dimer, with each monomer binding to one half-site to gain enough overall affinity for the promoter (Kuntz and Shapiro 1997). Homodimerization of liganded receptors is a nonlinear process because of the quadratic term describing the association between receptor monomers (Eqs (3) and (4)). In the absence of mixed-ligand heterodimer formation and at low doses of exogenous ligand X, newly formed XRRX cannot completely compensate for the loss of LRRL from receptor competition. As a result, even if X is a pure agonist or partial agonist with enough activity, X would initially reduce gene expression from the basal level instead of adding to it. At higher doses, the amount/activity of XRRX formed is able to completely replace and surpass lost LRRL, thereby reversing the downtrend in gene expression and producing a U-shaped dose-response curve. The essentiality of homodimerization to the occurrence of U-shape was demonstrated by linearization of this process, which produced
monotonic dose-responses in all circumstances. Although in our model receptor monomers form dimers prior to binding to HRE, a similar nonlinear response would also be expected if monomers were able to bind HRE sequentially and in a positively cooperative manner, forming the homodimer on the promoter.

As illustrated in Figure 4.2A, the magnitude of U-shape is positively correlated to endogenous hormone levels. The U-shape becomes less pronounced or would be too subtle to be identified if the endogenous hormone is at levels below its $K_d$ for the receptors ($K_d$ is 1 nM in the model). Given that endogenous hormones are usually at relatively low physiological levels, this may explain, at least in part, why SHR-mediated dose-responses are observed more often as monotonic rather than as U-shaped. Importantly, our results further showed that formation of mixed-ligand heterodimer LRRX can modulate the magnitude of U-shape observed with agonistic X. As the activity of LRRX becomes more influential, the overall shape of the dose-response curve for gene expression is increasingly amenable to the profile of LRRX, which itself is inverted U-shaped (Figure 4.4, bottom panels). If it is an activator, LRRX would first lessen the original U-shape of the dose-response curve, and then push it upward into an inverted U-shape; conversely if LRRX is a repressor, it would further deepen the original U-shape. In comparison, when the exogenous ligand X is an antagonist, no U-shape is observed, and only an inverted U-shape can be obtained when LRRX functions as a strong activator. Despite their significant role discussed here, it remains to be investigated whether LRRX indeed exist in cells in vitro and in vivo. But apparently, if they cannot be formed at all, cells would have a tendency to exhibit U-shaped responses when the exogenous ligand is an agonist.

Non-monotonic dose-responses in steroid or nuclear receptor signaling have been previously investigated through numerical simulations (Kohn and Portier 1993; Kohn and Melnick 2002; Conolly and Lutz 2004). Kohn and Portier had proposed that positive cooperative binding
between liganded receptor and DNA may result in a U-shaped dose-response curve (Kohn and Portier 1993). The origin of U-shape from the nonlinear dimerization process, as we noted in the present study, has a similar mathematical basis to their findings. Generation of U-shaped dose-responses with the latter mechanism relies, however, on the existence of more than one copy of the HRE in the promoter, which may not always be the case. With respect to the role of LRRX in U-shaped responses, Conolly and Lutz have relied on regarding them as transcriptionally inactive (Conolly and Lutz 2004) to explain the U-shaped response observed with AR agonist hydroxyflutamide (Maness, McDonnell et al. 1998). This transcriptionally inactive heterodimer is equivalent to the case of LRRX acting as a passive repressor in our model, which deepens the U-shape (Figure 4.5A). With respect to inverted U-shape, Kohn and Melnick suggested that one condition for this to occur is when there are excessive unoccupied receptors and recruitment of CoA by xenobiotic ligands is weaker than by endogenous ligands. However, in their model, receptor dimerization was not considered. In comparison, occurrence of inverted U-shaped curves in our model relies on the formation of LRRX functioning as activators. Additionally, as noted in the Method, an inverted U-shape can also arise if SHRs exist in such a high abundance that the excessive receptors may titrate away free CoA, provided DNA-independent association between these two species is allowed. This auto-inhibitory phenomenon due to self-squelching has been demonstrated in vitro (Bocquel, Kumar et al. 1989; Webb, Lopez et al. 1992).

SRMs represent a group of compounds whose activity, i.e., agonistic or antagonistic, varies in a cellular and tissue context-dependent manner. For example, tamoxifen and raloxifene are selective ER modulators, which are antiestrogenic in the breast but estrogenic in the bone (Dutertre and Smith 2000; Francucci, Romagni et al. 2005). Asoprisnil, a selective PR modulator, has an antiproliferative effect in primate endometrium, but can not induce labor in animal models of pregnancy and parturition (Chwalisz, Perez et al. 2005). Current understanding of the molecular mechanism for tissue-selective action of SRMs hinges on the notion that SRMs can
induce different conformational changes to their cognate SHRs, particularly in the C-terminal LBD, which in turn determines whether CoA or CoR will be recruited (Brzozowski, Pike et al. 1997; Shiau, Barstad et al. 1998). Conformational changes and differential coregulator recruitment also appear to be regulated by the phosphorylation status of SHRs and coregulators (Atanaskova, Keshamouni et al. 2002; Michalides, Griekspoor et al. 2004). For SRMs that are capable of recruiting, to some degree, both coactivators and corepressors, the relative abundance between these two types of opposing coregulators is a key to the direction of their genomic actions (Smith, Nawaz et al. 1997; Smith and O’Malley 2004; Wang, Blackford et al. 2004). In keeping with this concept, our simulation demonstrated that with a high CoR/CoA ratio, the activity of exogenous ligand X is primarily antagonistic, whereas with a low CoR/CoA ratio, the activity is primarily agonistic (Figure 4.9). Moreover, the present study revealed that the differential effect of SRMs could result from factors other than the CoR/CoA ratio. For instance, an exogenous ligand may have different affinities for the same type of receptor in different cell types or tissues. This affinity difference may cause, in a certain dose range, antagonistic activity in cells with lower affinity (resulting from the U-shape) and agonistic activity in cells with higher affinity (Figure 4.2B). In the presence of mixed-ligand heterodimer, the ability of LXXR to recruit CoA or CoR, which may depend on the phosphorylation status of the receptor and coactivator (Michalides, Griekspoor et al. 2004), can as well explain tissue selectivity within certain dose range (Figure 4.4B and 4.5B).

EACs represent a large set of environmental pollutants and naturally-occurring chemicals such as phytoestrogen that can interfere with the endocrine system. Many EACs act through SHRs to exert their endocrine disrupting effects (Amaral Mendes 2002; Markey, Rubin et al. 2002). Establishing and understanding the dose-response curves of EACs of interest constitutes an integral part of toxicological research and health risk assessment for these chemicals. The responses at low doses are particularly relevant to human health and may contribute to the
etiology of a variety of endocrine-related diseases (Dewailly, Ayotte et al. 1994; Lebel, Dodin et al. 1998; Snedeker 2001). Of interest, some EACs display biphasic effects within large dose ranges (Kemppainen and Wilson 1996; vom Saal, Timms et al. 1997; Maness, McDonnell et al. 1998; Calabrese 2001; Calabrese 2001; Putz, Schwartz et al. 2001; Putz, Schwartz et al. 2001; Almstrup, Fernandez et al. 2002; Terouanne, Nirde et al. 2003; Kohlerova and Skarda 2004). In parallel to this, the concept of hormesis has increasingly gained advocacy in recent years as a dose-response scheme for xenobiotics (Calabrese and Baldwin 2001; Calabrese 2005). A hormetic dose-response curve, either U- or inverted U-shaped, challenges the default linear model, which has been in practice for decades. A general explanation for hormetic phenomena has been that at low levels of disruption biological systems can launch compensatory responses that may overcorrect the initially perturbed state; while at higher doses the system may reach its maximum capacity for compensation, thus it is unable to counteract the perturbations (Calabrese 2004). Although quite common, this mechanism may not be the only explanation that can account for hormetic dose-response curves (Conolly and Lutz 2004; Weltje, vom Saal et al. 2005). The present study provides an additional yet distinct mechanism that may operate to produce U- and inverted U-shaped dose-responses to steroid mimics. Although perturbation of gene expression studied here only represents one of the initial responses in a series of molecular events leading to the adverse effects of EACs, the potential non-monotonic responses suggest that linear extrapolation for the low-dose effect of certain EACs may not be appropriate to evaluate their biological risks.

Experimentally obtained SHR-mediated dose–responses usually describe the averaged behavior of a population of cells. The ODE-based deterministic model presented in the present study was also designed to examine averaged responses. However, it is important to note that gene expression in individual cells is expected to be stochastic and may fluctuate to a great extent even at steady state (Thattai & van Oudenaarden 2001, Elowitz et al. 2002, Ozbudak et al. 2002,
Variations between cells in the abundance of receptors, coregulators, and other regulatory factors contribute further to the noise in gene expression as extrinsic sources. As a result, cells may respond differently to the same dose of an exogenous ligand. In certain cases, these heterogeneous responses may be critical to an individual cell’s fate decision as whether to proliferate, differentiate, or undergo apoptosis. An important implication of this effect is in breast cancer treatment with antiestrogen. Even though the majority of the cancer cells may respond to antiestrogenic therapy well, the heterogeneity arising from noisy gene expression may render a small fraction of the cancer cells insensitive to antiestrogen. These surviving cells may be responsible, at least in part, for the possibility of relapse of the disease after the termination of antiestrogenic therapy. Clearly, if individual cell behavior is important, a stochastic simulation approach is preferred. Nevertheless, the frequency of these cellular incidents in a population of isogenic cells in response to varying doses of exogenous ligands is otherwise deterministic, and may still be usefully modeled with an ODE-based approach. Another scenario is concerned with SHR controlling the gene expression of a secretory peptide hormone. Although in this situation individual cells may synthesize the hormone at very different levels due to gene expression noise, the overall output from the cell population, which is more relevant to the fitness of the organism as a whole, should be deterministic with respect to the dose of the exogenous ligand.

In conclusion, the present computational study revealed a novel mechanism, likely inherent to SHR-mediated steroid signaling, to explain the non-monotonic dose-responses and bidirectional effects observed with many steroid mimics. Our results may contribute to the understanding of how SRMs work and improve risk assessment for EACs.
4.6 Acknowledgements

We thank Dr. Kevin Gaido for his critical comments on the manuscript. This work is supported by funds from the Long-Range Research Initiative of the American Chemistry Council. The authors declare that there is no conflict of interest that would prejudice the impartiality of this scientific work.

4.7 References


Bledsoe RK, Stewart EL & Pearce KH 2004 Structure and function of the glucocorticoid receptor ligand binding domain. *Vitam Horm* 68 49-91.

Bocquel MT, Kumar V, Stricker C, Chambon P & Gronemeyer H 1989 The contribution of the N- and C-terminal regions of steroid receptors to activation of transcription is both receptor and cell-specific. *Nucleic Acids Res* 17 2581-2595.


Calabrese EJ 2004 Hormesis--basic, generalizable, central to toxicology and a method to improve the risk-assessment process. *Int J Occup Environ Health* 10 466-467.


Figure 4.1. Structure of the model for SHR-mediated gene expression. For detailed explanation of the signaling pathway, see Methods. For ordinary differential equations (ODEs) and choice of parameter values and references, see Table S1 and S2 in the Supporting Material. $k_{f_01}, k_{f_11}, k_{f_21}, k_{f_31}, k_{b_01}, k_{b_11}, k_{b_21}, k_{b_31}, k_{b_03}, k_{b_13}, k_{b_23}, k_{b_33}, k_{f_02}, k_{f_12}, k_{f_22}, k_{f_32}, k_{b_02}, k_{b_12}, k_{b_22}, k_{b_32}, k_{b_42}, k_{b_03}, k_{b_13}, k_{b_23}, k_{b_33}, k_{b_43}$ are association rate constants; $k_{b_01}, k_{b_11}, k_{b_21}, k_{b_31}, k_{b_02}, k_{b_12}, k_{b_22}, k_{b_32}, k_{b_42}, k_{b_03}, k_{b_13}, k_{b_23}, k_{b_33}, k_{b_43}$ are dissociation rate constants. Abbreviations: SHR: steroid hormone receptor; L: endogenous ligand; X: exogenous ligand; LR: L-bound SHR; XR: X-bound SHR; LRRL: homodimer of LR; XRRX: homodimer of XR; LRRX: mixed-ligand heterodimer of LR and XR; HRE: hormone response element; LRRLH: LRRL-bound HRE; XRRXH: XRRX-bound HRE; LRRXH: LRRX-bound HRE; CoA: coactivator; CoR: corepressor; CoALRRLH: LRRLH with CoA recruited; CoALRRXH: LRRXH with CoA recruited; CoAXRRXH: XRRXH with CoA recruited; CoRLRRXH: LRRXH with CoR recruited; CoRXRRXH: XRRXH with CoR recruited; GENEi: inactive gene; GENEa: active; PT: primary transcript; $\Phi$: degradation.
Figure 4.2. U-shaped steady-state dose-response in gene expression and magnitude of U-shape when exogenous ligand X is a pure agonist and the heterodimer LRRX is absent. (A) Effect of the level of endogenous ligand L. (B) Effect of binding affinity between X and SHR (implemented by varying the association rate constant $k_{f02}$). (C) Effect of SHR concentrations. (D) Effect of binding
affinity between XRs to form homodimer XRRX (implemented by varying the association rate constant $k_{12}$). Note: parameter values marked by * are default settings.

Figure 4.3. Linearization of the homodimerization process eliminates the U-shaped dose-response when exogenous ligand $X$ is a pure agonist and the heterodimer LRRX is absent. Refer to Equation 5 and 6 for method of linearization. (A) Effect of the level of endogenous ligand $L$. (B) Effect of binding affinity between $X$ and SHR (implemented by varying the association rate constant $k_{02}$). (C) Effect of SHR concentrations. (D) Effect of binding affinity between XRs to form homodimer XRRX (implemented by varying the association rate constant $k_{12}$).
**Figure 4.4.** Mixed-ligand heterodimer LRRX acting as an activator reduces and inverts the U-shaped dose-response when exogenous ligand X is a pure agonist. (A) Effect of binding affinity between LR and XR to form heterodimer LRRX (implemented by varying the association rate
Figure 4.5. Mixed-ligand heterodimer LRRX acting as a repressor deepens the U-shaped dose-response when exogenous ligand X is a pure agonist. (A) Effect of binding affinity between LR and XR to form heterodimer LRRX (implemented by varying the association rate constant $k_{f13}$). In this case, LRRX was simulated as a passive repressor by setting $k_{f23}$ to zero. (B) Effect of the ability of LRRX to recruit CoR (implemented by varying the association rate constant $k_{f43}$).
Figure 4.6. Monotonically decreasing dose-response in gene expression when exogenous ligand X is a pure antagonist and the heterodimer LRRX is absent. (A) Effect of binding affinity between X and SHR (implemented by varying the association rate constant $k_{f02}$). In this case, X was simulated as a passive antagonist by setting $k_{f12}$ to zero. (B) Effect of binding affinity between XRRs to form homodimer XRRX (implemented by varying the association rate constant $k_{f12}$). In this case, X was simulated as a passive antagonist by setting $k_{f22}$ to zero. (C) Effect of binding affinity between XRRX and HRE (implemented by varying the association rate constant $k_{f22}$). In this case, X was simulated as a passive antagonist by setting $k_{f42}$ to zero. (D) Effect of the ability of XRRX to recruit CoR (implemented by varying the association rate constant $k_{f42}$).
Figure 4.7. Mixed-ligand heterodimer LRRX acting as an activator rightward shifts and converts to inverted U-shape, the monotonically decreasing dose-response when exogenous ligand X is a pure antagonist. (A) Effect of binding affinity between LR and XR to form heterodimer LRRX (implemented by varying the association rate constant $k_{f13}$). (B) Effect of the ability of LRRX to recruit CoA (implemented by varying the association rate constant $k_{f33}$).
Figure 4.8. Mixed-ligand heterodimer LRRX acting as a repressor leftward shifts the monotonically decreasing dose-response when exogenous ligand X is a pure antagonist. (A) Effect of binding affinity between LR and XR to form heterodimer LRRX (implemented by varying the association rate constant $k_{f13}$). In this case, LRRX was simulated as a passive repressor by setting $k_{f23}$ to zero. (B) Effect of the ability of LRRX to recruit CoR (implemented by varying the association rate constant $k_{f43}$).
Figure 4.9. Effect of CoR/CoA ratio on the dose-response when exogenous ligand X is a partial agonist. To keep the basal expression level sustained by L unchanged, only the initial
concentration of CoR was varied to achieve different CoR/CoA ratios. **(A)** In the absence of heterodimer LRRX. **(B)** LRRX acting as an active repressor. **(C)** LRRX acting as a pure activator. In this case, $k_{r13}$ was set to 50% of the default value in the left panel, and 2-fold in the right panel. **(D)** LRRX acting as a partial activator. In this case, $k_{r13}$ was set to 50% of the default value in the left panel, and 4-fold in the right panel.

**Table 4.1.** Conditions in which U- or inverted U-shaped dose-response curves may appear.

<table>
<thead>
<tr>
<th>Exogenous ligand X</th>
<th>Mixed-ligand Heterodimer LRRX</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>Pure or partial activator</td>
</tr>
<tr>
<td>Pure or partial agonist</td>
<td>U-shape</td>
<td>U-shape, inverted U-shape</td>
</tr>
<tr>
<td>Antagonist</td>
<td>none</td>
<td>Inverted U-shape</td>
</tr>
</tbody>
</table>

Note: - indicates LRRX is absent; + indicates LRRX is present.
4.8 Supporting Material

Table 4.S1. Species, initial values, and ODEs for the SHR-mediated gene expression model.

<table>
<thead>
<tr>
<th>Species</th>
<th>Default initial value</th>
<th>Ordinary differential equations</th>
</tr>
</thead>
</table>
| L       | 60                    | \[
\frac{d[\text{L}]}{dt} = 0
\] |
| X       | varied                | \[
\frac{d[\text{X}]}{dt} = 0
\] |
| SHR     | 120                   | \[
\frac{d[\text{SHR}]}{dt} = -k_{f_{01}}[\text{SHR}][L] + k_{b_{01}}[\text{LR}] - k_{f_{02}}[\text{SHR}][X] + k_{b_{02}}[\text{XR}]
\] |
| LR      | 0                     | \[
\frac{d[\text{LR}]}{dt} = k_{f_{11}}[\text{SHR}][L] - k_{b_{01}}[\text{LR}] - 2k_{f_{11}}[\text{LR}]^2 + 2k_{b_{11}}[\text{LRLL}] - k_{f_{13}}[\text{XR}][LR] + k_{b_{11}}[\text{LRXX}]
\] |
| XR      | 0                     | \[
\frac{d[\text{XR}]}{dt} = k_{f_{02}}[\text{SHR}][X] - k_{b_{02}}[\text{XR}] - 2k_{f_{12}}[\text{XR}]^2 + 2k_{b_{12}}[\text{LRXX}][LR] - k_{f_{13}}[\text{XR}][LR] + k_{b_{13}}[\text{LRXX}]
\] |
| LRRL    | 0                     | \[
\frac{d[\text{LRRL}]}{dt} = -k_{f_{11}}[\text{LR}]^2 - k_{b_{11}}[\text{LRRL}]
\] |
| LRRLH   | 0                     | \[
\frac{d[\text{LRRLH}]}{dt} = k_{f_{21}}[\text{LRRL}][\text{HRE}] + k_{b_{21}}[\text{LRRLH}]
\] |
| XRRX    | 0                     | \[
\frac{d[\text{XRRX}]}{dt} = k_{f_{12}}[\text{XR}]^2 - k_{b_{12}}[\text{XRRX}]
\] |
| LRRX    | 0                     | \[
\frac{d[\text{LRRX}]}{dt} = k_{f_{13}}[\text{XR}][\text{LR}] - k_{b_{13}}[\text{LRRX}]
\] |
| HRE     | 1                     | \[
\frac{d[\text{HRE}]}{dt} = -k_{f_{21}}[\text{LRRL}][HRE] + k_{b_{21}}[\text{LRRLH}]
\] |
| LRRXLH  | 0                     | \[
\frac{d[\text{LRRXLH}]}{dt} = k_{f_{21}}[\text{LRRL}][HRE] - k_{b_{21}}[\text{LRRLH}]
\] |
| XRRXH   | 0                     | \[
\frac{d[\text{XRRXH}]}{dt} = k_{f_{22}}[\text{XRRX}][\text{HRE}] - k_{b_{22}}[\text{XRRXH}]
\] |
<table>
<thead>
<tr>
<th>Species</th>
<th>Default initial value</th>
<th>Ordinary differential equations</th>
</tr>
</thead>
</table>
| LRRXH          | 0                     | \[
\begin{align*}
\frac{d[LRRXH]}{dt} &= k_{f_{123}}[LRRX][HRE] - k_{b_{23}}[LRRXH] \\
&- k_{f_{33}}[LRRXH][CoA] + k_{b_{33}}[CoALRRXH] \\
&- k_{f_{43}}[LRRXH][CoR] + k_{b_{43}}[CoRLRRXH]
\end{align*}
\] |
| CoA            | 60                    | \[
\begin{align*}
\frac{d[CoA]}{dt} &= -k_{f_{31}}[LRLHL][CoA] + k_{b_{31}}[CoALRLHL] \\
&- k_{f_{32}}[XRRXH][CoA] + k_{b_{32}}[CoAXRRXH] \\
&- k_{f_{33}}[LRRXH][CoA] + k_{b_{33}}[CoALRRXH]
\end{align*}
\] |
| CoR            | 60                    | \[
\begin{align*}
\frac{d[CoR]}{dt} &= -k_{f_{42}}[XRRXH][CoR] + k_{b_{42}}[CoRXRRXH] \\
&- k_{f_{43}}[LRRXH][CoR] + k_{b_{43}}[CoRLRRXH]
\end{align*}
\] |
| CoALRRHLH      | 0                     | \[
\begin{align*}
\frac{d[CoALRLHL]}{dt} &= k_{f_{31}}[LRLHL][CoA] - k_{b_{31}}[CoALRLHL]
\end{align*}
\] |
| CoAXRRXH       | 0                     | \[
\begin{align*}
\frac{d[CoAXRRXH]}{dt} &= k_{f_{32}}[XRRXH][CoA] - k_{b_{32}}[CoAXRRXH]
\end{align*}
\] |
| CoALRRXH       | 0                     | \[
\begin{align*}
\frac{d[CoALRRXH]}{dt} &= k_{f_{33}}[LRRXH][CoA] - k_{b_{33}}[CoALRRXH]
\end{align*}
\] |
| CoRXRRXH       | 0                     | \[
\begin{align*}
\frac{d[CoRXRRXH]}{dt} &= k_{f_{42}}[XRRXH][CoR] - k_{b_{42}}[CoRXRRXH]
\end{align*}
\] |
| CoRLRRXH       | 0                     | \[
\begin{align*}
\frac{d[CoRLRRXH]}{dt} &= k_{f_{43}}[LRRXH][CoR] - k_{b_{43}}[CoRLRRXH]
\end{align*}
\] |
| GENE_i         | 1                     | \[
\begin{align*}
\frac{d[GENE_i]}{dt} &= -(k_{f_{51}}[CoALRRH] + k_{f_{52}}[CoAXRRXH] + k_{f_{53}}[CoALRRXH]) \cdot [GENE_i] \\
&+ (k_{b_{51}} + k_{b_{52}})[CoRXRRXH] + k_{b_{53}}[CoRLRRXH]) \cdot [GENE_a]
\end{align*}
\] |
| GENE_a         | 0                     | \[
\begin{align*}
\frac{d[GENE_a]}{dt} &= (k_{f_{51}}[CoALRRH] + k_{f_{52}}[CoAXRRXH] + k_{f_{53}}[CoALRRXH]) \cdot [GENE_i] \\
&- (k_{b_{51}} + k_{b_{52}})[CoRXRRXH] + k_{b_{53}}[CoRLRRXH]) \cdot [GENE_a]
\end{align*}
\] |
| mRNA           | 0                     | \[
\begin{align*}
\frac{d[mRNA]}{dt} &= k_{f_{8}}[PT] - k_{f_{9}}[mRNA]
\end{align*}
\] |
| Protein        | 0                     | \[
\begin{align*}
\frac{d[protein]}{dt} &= k_{f_{10}}[mRNA] - k_{f_{11}}[protein]
\end{align*}
\] |

**Note:** Initial values represent the copy number of species. With the assumption that the nuclear volume is 100 $\mu$m$^3$, the equivalent molar concentrations for L, SHR, CoA, and CoR are 1, 2, 1, and 1 nM, respectively. These values are compatible with the concept that transcription factors and coregulators usually exist in low concentrations.
abundance. The reported concentration of SHRs is generally in the nanomolar range (Furlow, Murdoch et al. 1993; Rossini and Malaguti 1994).
Table 4.S2. Default reaction constant values for the SHR-mediated gene expression model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default value (s(^{-1}))</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k_{01}, k_{02})</td>
<td>1.66x10(^{-5})</td>
<td>When converted back to molar concentration, the association rate constant 1.66x10(^{-5})/s is equivalent to 1x10(^{-3})/nM/s, which is close to the value of 1.3x and 1.61x10(^{-3})/nM/s measured between 17(\beta)-estradiol and ER in two separate studies (Weichman and Notides 1977; Rich, Hoth et al. 2002). The dissociation rate constant 1x10(^{-3})/s is close to the value of 1.2x and 2x10(^{-3})/s reported in the same two studies. The pair of values used here gives a Kd of 1nM.</td>
</tr>
<tr>
<td>(k_{001}, k_{002})</td>
<td>1x10(^{-3})</td>
<td></td>
</tr>
<tr>
<td>(k_{11}, k_{12}, k_{13})</td>
<td>3.0x10(^{-7})</td>
<td>Reported Kd between SHR monomers such as ER ranges from low nM to around 50 nM (Brandt and Vickery 1997; Tamrazi, Carlson et al. 2002; Jisa and Jungbauer 2003). The pair of values used here gives a Kd of 12 nM. The dissociation rate constant was set according to the average obtained for ER dimers (Jisa and Jungbauer 2003).</td>
</tr>
<tr>
<td>(k_{011}, k_{012}, k_{013})</td>
<td>2.2x10(^{-4})</td>
<td></td>
</tr>
<tr>
<td>(k_{21}, k_{22}, k_{23})</td>
<td>1.16x10(^{-4})</td>
<td>These values are derived from binding kinetics measured between estradiol-liganded ER dimer and ERE (Ozers, Hill et al. 1997; Szatkowski Ozers, Hill et al. 2001). This pair of values gives a Kd close to 2nM. AR dimer and ARE were reported to have a Kd of 2 nM (Liao, Zhou et al. 1999).</td>
</tr>
<tr>
<td>(k_{021}, k_{022}, k_{023})</td>
<td>1.48x10(^{-2})</td>
<td></td>
</tr>
<tr>
<td>(k_{31}, k_{32}, k_{33})</td>
<td>2.32x10(^{-5})</td>
<td>Reported Kd between ER dimer and various coactivators ranges from low nM to several hundred nM (Suen, Berrodin et al. 1998; Thenot, Bonnet et al. 1999; Tikkanen, Carter et al. 2000; Margeat, Poujol et al. 2001; Warnmark, Almlof et al. 2001; Wong, Komm et al. 2001; Warnmark, Treuter et al. 2002; Cheskis, McKenna et al. 2003). The pair of values used here gives a Kd of 10 nM. The dissociation rate constant was set the same as that for the receptor dimer and HRE complex, which gives an average lifetime of 67s, compatible with the rapid exchange observed between coregulators and transcription factors (Stenoien, Nye et al. 2001; Becker, Baumann et al. 2002).</td>
</tr>
<tr>
<td>(k_{031}, k_{032}, k_{033})</td>
<td>1.48x10(^{-2})</td>
<td></td>
</tr>
<tr>
<td>(k_{42}, k_{43})</td>
<td>2.32x10(^{-5})</td>
<td>This pair of default values was assumed the same as that for coactivators.</td>
</tr>
<tr>
<td>(k_{442}, k_{443})</td>
<td>1.48x10(^{-2})</td>
<td></td>
</tr>
<tr>
<td>(k_{51})</td>
<td>5.56x10(^{-4})</td>
<td>Chromatin decondensation and condensation, representing gene activation and deactivation, is a slow process (Muller, Walker et al. 2001). The value of (k_{51}) gives an average GENE(_{a}) lifetime of 30 min. The values of other parameters were assumed.</td>
</tr>
<tr>
<td>(k_{52}, k_{53})</td>
<td>3.34x10(^{-2})</td>
<td></td>
</tr>
<tr>
<td>(k_{51}, k_{52}, k_{53})</td>
<td>1.67x10(^{-3})</td>
<td></td>
</tr>
<tr>
<td>(k_{6})</td>
<td>2.78x10(^{-3})</td>
<td>With this value, about 10 primary transcripts (PTs) are produced per gene template per hour.</td>
</tr>
<tr>
<td>Parameter</td>
<td>Default value (s⁻¹)</td>
<td>Notes</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------</td>
<td>-------</td>
</tr>
<tr>
<td>k₇</td>
<td>3.34x10⁻³</td>
<td>The ratio of k₇/k₈ was set to 2 to reflect the fact that only about 1/3 of PTs become mature mRNA and the rest are degraded within the nucleus (Jackson, Pombo et al. 2000). Additionally, it takes about 10–20 min for PTs to mature and translocate to the cytoplasm as mRNA (Jackson, Pombo et al. 2000). The value for k₈ used here gives an average 10 min for this process, and k₇ was then set according to the ratio.</td>
</tr>
<tr>
<td>k₈</td>
<td>1.67x10⁻³</td>
<td></td>
</tr>
<tr>
<td>k₉</td>
<td>3.2x10⁻⁵</td>
<td>Half-life values for luciferase, a common reporter gene, were used. This value gives an mRNA half-life of 6 h. It was numerically derived based on a study in which luciferase mRNA was delivered to B16-F10 cells (Bettinger, Carlisle et al. 2001). This half-life is also about the same as Promega (Madison, WI) provided for firefly luciferase mRNA.</td>
</tr>
<tr>
<td>k₁₀</td>
<td>4.17x10⁻²</td>
<td>This value gives a translation rate of 150 protein molecules per mRNA template per hour, an average found in eukaryotic cells (Jackson, Pombo et al. 2000; Arava, Wang et al. 2003).</td>
</tr>
<tr>
<td>k₂₀</td>
<td>6.4x10⁻⁵</td>
<td>This value gives a protein half-life of 3 h, which is in the range of 50 min to 3.68 h for firefly luciferase in mammalian cells (Thompson, Hayes et al. 1991; Nunez, Faught et al. 1998; Leclerc, Boockfor et al. 2000; Ignowski and Schaffer 2004).</td>
</tr>
</tbody>
</table>
4.10 Support Material References


Chapter 5

Conclusions
The completion of the Human Genome Project is leading biological and medical research into a new information science, capable of taking global views of biological systems. A large amount of high-throughput genetic and proteomic data provide opportunities for new research targets and biomedical applications. The focus of investigation therefore has been transformed from data accumulation into gene regulation, function, and molecular interactions. However, transformation of the sequence data into useful biological information is a challenging task. One of the main tasks of the informatics science is to systematically investigate all molecules and their interactions within living cells so as to understand how these molecules and the interactions between them relate to the function of the organism as a whole (Zhao 2006). Full understanding of biological networks requires learning the constituting elements and unraveling the dynamic interactions between genes, transcripts, proteins and pathways.

Androgen receptors (AR) are the critical factors responsible for the development of the male phenotype during embryogenesis and for the achievement of sexual maturation at puberty (Heemers 2006). Typically, AR is known to exert its activity by binding to androgen response elements (ARE) located in the promoter region of target genes, thereby directly affecting the expression of these genes. This regulation may be defined as a primary response. Regulation of gene expression may also be achieved through a secondary response via regulatory proteins that are induced by androgen in the primary response. The cross-talk between intracellular receptors and other signal transduction pathways may be responsible for the regulation of secondary response genes (Dean 1996). A large amount of evidence indicates that androgens may indirectly influence the expression of genes lacking AREs by modulating the activity of secondary transcription factors, mediating the expression of growth factors, or by inducing changes in the production of other hormones (Heemers 2006). These sequential effects of androgens can induce cascade-like actions and may play an important role in more complex processes involving coordinated responses of genes, cells, and organs. To systematically elucidate androgen
signaling, we describe five possible scenarios of androgen responses through genomic and non-genomic pathways (Figure 5.1). In the first scenario, the genomic effects require that androgens passively diffuse into cells to act as a transcription factor after binding to its nuclear receptor. Primary response genes are then transcriptionally activated or suppressed after binding of ligand-activated AR to AREs in their 5’-flanking regulatory regions (Zhou 2005). In addition, non-genomic effects may occur through AR targeted to the plasma membrane or through a membrane-associated receptor, which in turn activates other transcription factors (TF), as shown in scenarios 2 and 3. Androgen therefore binds to AR as well as the membrane receptor, where

Figure 5.1. Schematic representation for the mechanisms of action of androgen. Scenario 1, the direct interaction of AR complex with the regulatory DNA of the target primary response gene; Scenario 2 and 3, non-genomic actions of androgen through a membrane receptor; Scenario 4, one of the mechanisms by which androgen could affect the induction of a regulatory protein that ultimate affects the transcription of the secondary response genes; Scenario 5, one of the mechanisms whereby transcription of a primary response gene could be delayed, resulting in a gene falling into the delayed primary response gene category. AR: androgen receptor; TF: other transcription factors. TF$_\text{early}$: Early gene products induced by androgens, which are required for transcription of the secondary response genes.
the membrane receptor activates a transcription factor through alternative signaling transduction pathways, such as MAPK. Subsequently, the androgen receptor could associate with other transcription factors activated by a membrane receptor, in turn binding to target genes, triggering transcription (Figure 5.1, scenario 2). For instance, biological evidence suggests AR and glucocorticoid receptor (GR) indeed interact at the transcriptional level by forming heterodimers (Chen 1997). However, the third scenario entails only the binding of androgen to a membrane receptor, activating a transcription factor, and then promoting gene transcription. These first three scenarios encompass early gene transcription. In contrast, the fourth scenario involves delayed response genes responding to androgen stimulation via AR binding. Although these secondary response genes contain ARE motifs, the response requires interaction with other transcription factors and synthesis of concomitant proteins from the early stage. Thus there are genes that apparently bind AR directly, yet exhibit a delayed induction. Finally, secondary response genes without AREs also can respond indirectly to androgen stimulation via activation or suppression of transcription by other ongoing factors directly regulated by AR (Figure 5.1, scenario 5). This scenario comprises the broad network of androgen regulated early gene products binding to androgen-mediated target genes, thereby promoting transcription. The responses in the fourth and fifth scenarios require the synthesis of new transcription factor molecules, therefore they have a temporal delay.

After a brief discussion of the present understanding of the general mechanism of gene regulation by androgens, we are interested in establishing androgen-regulated transcriptional networks and distinguish between different scenarios of androgen action. Perhaps the most important pathway is that mediated via AR directly which controls expression of androgen-regulated genes (Figure 5.1, scenarios 1 and 4). In the case of primary response genes, the intracellular ARs act as transcription factors and directly bind, or indirectly bind through tethering, to AREs within the
target genes (Dean 1996). To group genes based upon the mechanisms through which they are induced by intracellular receptors, three statistical methods, Hidden Markov Model (HMM), Support Vector Machine (SVM), and Random Forest (RF) were explored to strengthen the identification of putative AREs in the human genome. The investigation in Chapter 2 was based on specific models derived from experimentally verified known AREs. Although these models were built on different features of the same training set, cross-validation results indicated that all three methods provided good sensitivity and specificity in identifying AREs, with an accuracy rate of at least 80%. Applied to the –5,000bp upstream promoter regions of the human genome, the HMM approach identified around 390 AR binding sites out of 23,570 genes analyzed. However, the other two approaches defined approximately 1,000 AREs, as the number of sites was larger than the expected frequency in random DNA sequences. The number of random AREs was calculated from the total number of base pairs in the genome divided by the frequency of a sequence with specified based pairs at 12 positions, which leads to 192 sites (Horie-Inoue 2004). Also, in a recent paper (Bolton 2007), Bolton et al. performed chromatin immunoprecipitation on human prostate epithelial cell line to identify 524 AR binding regions and validated them in reporter assays. Interestingly, 67% of their AREs resided within -50 kb of the transcription start sites. Therefore, the number of ARE genes identified by the SVM or RF models is far greater than the expected number of real sites. Differences in feature selection between SVM, RF and HMM may contribute to discrepancies between the search methods. More false positive AREs identified by SVM and RF models is unsurprising because the sequence features extracted are likely to contain other transcription factor binding sites and promoters that are not specific to AREs. In addition, the locations of ARE motifs were known from experiments, so we selected features based on the narrow 15 bp of core AREs and the 100bp of flanking region. Without knowing where AREs are located in any testing sequences, features were generated from a much larger 5000 bp region, not necessarily representative of the training data set. One fundamental complication is that the training data was based on a small region of the ARE-containing genes,
while the testing model used a much larger DNA region. It therefore follows that more false positive sites will be predicted using SVM and RF models. We hypothesize that by adding more relevant features, the accuracy of SVM and RF prediction could be improved in future studies. However, since prediction results are largely dependent on the available resources, difficulties exist in obtaining experimental data for ARE-specific features. Nevertheless, the HMM model offers a systematic approach to predicting genes directly regulated by AR on the basis of their promoter sequences so the HMM model could be employed toward identifying positive ARE candidates. The prediction of genes containing AREs is a basic step for the understanding of the mechanisms of AR transcriptional regulatory gene networks, and those genes may represent the most promising candidates for further functional analysis.

Understanding the underlying mechanism and dynamics of AR regulated gene networks requires knowing not only the direct target genes modulated by the androgen receptor signaling pathway,

![Venn diagrams](image)

**Figure 5.2.** Venn diagrams summarize the number of differentially expressed genes at different time points. **A.** Late and early responses for ARE positive and negative genes were compared and the numbers of common genes are shown. Late and early responses for the subsets of **B.** ARE positive genes only and **C.** ARE negative genes only are also given.
but also the interaction with other transcription regulators which are shown in scenarios 2, 3 and 5 (Figure 5.1). The elucidation of networks for modulating AR activity in cells certainly requires genome-wide approaches such as microarray analysis. In Chapter 3, we have pursued the existence of common regulatory motifs in AR-regulated genes identified by microarray analysis. First, we analyzed the gene expression profile of human LNCaP cell line treated for 6 and 24 hours with exogenous androgen. Analyzed by GeneSpring 7, the result indicated that there were a total of 745 significantly expressed genes. Of 745 putative androgen responsive genes identified, 503 were induced and 242 were repressed. To identify regulatory elements that control AR-mediated gene expression patterns, sequence-based and motif-based analyses in the upstream regions of differentially expressed gene groups were performed. The application of promoter classification linked known transcription factors into a model of AR-mediated gene expression and revealed transcriptional control mechanisms. Promoters that have been characterized in AR-mediated gene expression are those for C/EBP, CREB, PAX, and GATA factors (Table 5.1). Genes with those binding sites play important roles in primary, cellular metabolic processes and are involved in cell proliferation and differentiation. Deciphering the cross talk between biomechanical interactions between androgen, the plasma membrane and AR is fundamental to a clear understanding of androgen-dependent effects. Furthermore, we identified gene sets that could distinguish between early (Figure 5.1, scenarios 1, 2, 3) and late responses (Figure 5.1, scenarios 4, 5). A Venn diagram analysis presents the distribution of genes identified at early and late stages (Figure. 5.2A). The total number of differentially expressed genes at each time point does not differ significantly. Among these, 84 genes were in common between early and late gene lists. The occurrence of early and late effects indicates that there are at least two kinds of androgen responses. Therefore, we performed a search of ARE sites for genes directly regulated by AR among early and late genes. The results of our analysis are shown in Figures 5.2B and 5.2C. These diagrams provide an overview of the existence of overlap between early and late gene groups for positive AREs (Figure 5.2B) and negative AREs
(Figure 5.2C) respectively. Given 125 genes and 108 genes involved in genomic AR and in the secondary response, the analysis could discriminate between scenario 1 and scenario 4. To discover motifs that occur among early and late gene groups, MotifScanner software was used to screen common transcription factor motifs in positive ARE and negative ARE gene groups at early and late stages. The distribution of transcription factors among different groups is shown in histograms form (Figure 5.3 and 5.4). The results indicated that the frequencies of the TFBS are similar in early and late genes with p-values greater than 0.05, as shown in a table (Figure 5.3 and 5.4). The AR-mediated gene expression time course analysis indicated few significant

Table 5.1. After scanning upstream sequences for AR- responsive and nonresponsive genes using MotifScanner, the most discriminative TFBSs between response and non-response groups were determined. The table shows top ranked TFBS names and recognition sequences, along with their corresponding motifs.

<table>
<thead>
<tr>
<th>Motif</th>
<th>Sequences</th>
<th>TFBSs ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAX</td>
<td>GAAAAAATAAC CAAAATTAGC</td>
<td>M00377-V$PAX4_02</td>
</tr>
<tr>
<td>EBOX</td>
<td>CTTCACCTGGT TTTCACCTTG</td>
<td>M00073-V$DELTAEF1_01</td>
</tr>
<tr>
<td>Fork Head Domain factors</td>
<td>TCTGTGAATAAAATA TATGTAAAAAAACT</td>
<td>M00269-V$XFD3_01</td>
</tr>
<tr>
<td>Myeloid Transforming protein</td>
<td>TGAGAAGATTTAATAA GGAGAAGCCTAGATA</td>
<td>M00078-V$EVI1_01</td>
</tr>
<tr>
<td>Caudal related homeodomain</td>
<td>CTTTATA ATTTATA</td>
<td>M00100-V$CDXA_01</td>
</tr>
<tr>
<td>Myogenin / nuclear factor 1</td>
<td>ATCTTGAATAAAATT</td>
<td>M00190-V$CEBP_Q2</td>
</tr>
<tr>
<td>or related factors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CREB</td>
<td>TGGCGCTGACGACGA AATGGCTGACGTCTA</td>
<td>M00114-V$TAXCREB_01</td>
</tr>
<tr>
<td>GATA</td>
<td>TCATTGATAATAGA</td>
<td>M00126-V$GATA1_02</td>
</tr>
<tr>
<td>Myogenin / nuclear factor 1</td>
<td>CACTGCTTTTAGGTCTGCCTGCC TGACTGGTTTTGGCACAAAACTCAGCA</td>
<td>M00056-V$MYOGNF1_01</td>
</tr>
<tr>
<td>or related factors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CREB</td>
<td>TGACGTAA TGACATAA</td>
<td>M00040-V$CREBP1_01</td>
</tr>
</tbody>
</table>
Figure 5.3. Distribution of transcription factor binding sites (TFBS) common in ARE positive gene groups for early responsive genes, the overlap between early and late responsive genes, and late responsive genes. Table shows TFBS and corresponding motif names. Data was normalized to account for different group sizes. p-value was calculated for each individual TFBS among three different groups using chi-square test.
**Figure 5.4.** Distribution of transcription factor binding sites common in ARE negative gene groups for early responsive genes, the overlap between early and late responsive genes, and late responsive genes. Table shows TFBS and corresponding motif names. Data was normalized to account for different group sizes. p-value was calculated for each individual TFBS among three different groups using chi-square test.
changes between early and late gene response. Microarray data is expected to contribute significantly to the overall picture, necessitating the integration of genes and pathways, one of the basic goals of systems biology (Werner 2007). In higher eukaryotes, regulatory elements tend to be short and dispersed among long intergenic sequences, and their identification is difficult (Liu 2004). All of the existing prediction methods require that the known binding sites be represented as a consensus of the collection or as a matrix of acceptable nucleotides at each position (Elnitski 2006). These methods are convenient but may exclude a significant subset of true binding site elements because of omission of important variable regions. This may restrict us from finding real TFBSs due to the limited amount of training data available. Developing new prediction algorithms is necessary to aid regulatory element identification.

The availability of microarrays, coupled with genome sequence information, yields useful information about transcription control and regulatory networks. However, the results are static observations which do not directly reveal the underlying mechanisms that represent the real link to transcription factors and their actions. A fundamental tool for elucidating the underlying patterns of interaction to obtain more realistic and predictive relationships in a cell is the dynamically modeled gene regulatory network. It is clear that the parts list in the networks is far from complete, as is the connection of genes to basic cellular roles. Instead of simulating entire organisms, we can begin by dynamically modeling a sub-network or a small module (Roth 2007). Fortunately, differential equation-based computational models, exciting biomedical approaches in the examination of signaling processes, are of value in facilitating the understanding of many conditions. On the path towards simulating subsystems of AR-mediated gene expression, we assembled knowledge about individual physical interactions to model genomic responses to exogenous androgen as shown in scenario 1 (Figure 5.1). Computational analysis of interacting cellular networks appears most likely to provide the basic information needed to draw mechanistic inferences of the shapes of dose–response curves (Andersen 2005). In Chapter 4, an Ordinary
Differential Equation (ODE)-based computational model was constructed to investigate dose-response behavior within the simple framework of classical steroid hormone receptor signaling. With this model, we varied the parameters and their connections to determine what role each plays in network function. Simulation results found that non-monotonic dose-response curves resulted under various conditions. Specifically, when the exogenous ligand is an agonist, a U-shaped dose–response appears as a result of the inherently nonlinear process of receptor homodimerization. Additionally, we found that an inverted U-shaped dose–response can arise when the heterodimer is a strong transcriptional activator regardless of whether the exogenous ligand is an agonist or antagonist. Modeling studies allow us to determine the role that each reactant plays in shaping different dose curves, enabling us to determine the consequence of network outputs. In summary, the computational study in Chapter 4 revealed a novel mechanism, likely inherent to SHR-mediated steroid signaling, to explain the non-monotonic dose-responses and bidirectional effects observed with many steroid mimics. As an example of dynamic subsystem modeling, the model of AR-regulated gene expression has proven useful in understanding molecular mechanisms underlying biological systems. Nevertheless, to simulate entire tissues or organisms, a comprehensive predictive model of androgen-modulated gene networks will be built when a complete understanding of the relationships between components of AR regulatory networks are available.

The approaches used so far contribute to establishing the topology of AR transcriptional networks and the underlying mechanism of their regulatory connections. Although much still remains to be understood about the molecular pathways involved in androgen regulation, it is generally accepted that transcriptional elements control gene expression patterns. Based on this premise, we adopted a supervised learning tool to discriminate between genes involved in early or late androgen responses. However, the results indicated no significant TFBS between the two groups. The supervised method is convenient but may exclude a significant subset of a binding site
repertoire because of omission of important variable regions. To achieve more accurate results, unsupervised learning approaches could be constructed to identify TFBS among different androgen responsive genes (ARG) in a future study. Unsupervised methods, such as hierarchical clustering are thought to reveal some set of common but unknown patterns, thus leading to discoveries of similarities in each cluster. In addition, we may consider gene function annotations in Gene Ontology (GO) as one feature for classification. It is reasonable to hypothesize that a set of coexpressed genes will share the same functions. After obtaining a subset of AR-regulated genes along with their respective binding predictions, we will collect the associated GO terms, and construct a set of gene functions that are overrepresented in known androgen targets as represented in the GO terms. By incorporating nonexpression-based data sources, we believe that the accuracy of prediction could be improved.

Furthermore, there are several published ODE-based MAPK computational models (Huang 1996; Bhalla 2002; Reynolds 2003), which could be linked with our AR-mediated gene expression module. A significant body of literature shows that phosphorylation of AR and its coregulators by MAPK appears to play an important role in integrating membrane-initiated effects (Yang 2003). To expand our dynamic model and focus on non-genomic action effects on dose response behaviors, MAPK models could be integrated into our model for AR-mediated transcription. Although there is no complete map connecting multiple signaling kinase pathways and androgen-mediated transcription, the integrated model would provide new insight into mechanisms of different shapes of dose-response curves regulated by androgens and improve prediction of EAC exposure health risks. In addition, our dynamic AR signaling model is based on differential equations, which inevitably have numerous unknown parameters. An attractive alternative is given by a Boolean approach, which allows cyclic behavior in temporal expression pattern, while also quantifying the influence of one gene on another (Soneji 2007). Given genes involved in androgen-mediated signaling from microarray analysis, we will continue to build interactive
pathways representing the key genes and their products using Ingenuity Pathway Analysis (www.ingenuity.com). Based on a binary ON/OFF representation of mRNA and protein levels (Andersen 2005), the Boolean approach can readily be applied to androgen-mediated gene regulatory networks and to reproduce gene expression pattern. With a Boolean model, we can gain insight into gene interaction by studying the group of genes that can be activated by the expression of a given gene. This will be an important first step in understanding the interplay between the topology and functioning of androgen-mediated gene regulatory networks (Albert 2003), and it could provide a simple verification of the sufficiency of gene interactions.

In conclusion, the recent years have exuded great promise in the understanding of gene networks. We are now at a time where complicated multifaceted relationships can begin to be unraveled by the high caliber of sophisticated computational approaches. It is our hope that other researchers will build upon our methods to further the science of gene regulatory networks.
5.1 References:


