MERJA MATILAINEN

Identification and Characterization of Target Genes of the Nuclear Receptors VDR and PPARs

Implementing in silico Methods Into the Analysis of Nuclear Receptor Regulomes

Doctoral dissertation

To be presented by permission of the Faculty of Natural and Environmental Sciences of the University of Kuopio for public examination in Auditorium L22, Snellmannia building, University of Kuopio, on Saturday 17th March 2007, at 12 noon

Department of Biosciences
University of Kuopio
ISSN 1235-0486

Abstract

Homeostasis in humans is maintained via an integrated regulation of gene expression, for which the first and crucial step is the control of their transcription rate. Ligand-activated nuclear receptors (NRs) are key transcription factors (TFs) that translate both nutritional and hormonal signals into changes in gene expression. Structurally, these TFs have a ligand-binding domain (LBD), the structural features of which affect the transactivation events that lead to change in target gene expression. The link to DNA is formed via the DNA-binding domain (DBD) that recognizes specific DNA-binding elements, called response elements (REs). Prediction of NR target genes relies on the identification of putative NR REs in the regulatory regions of genes.

The main objective of the present study was to implement bioinformatics in silico methods into the characterization of structural features that underlie the functional differences seen in the NRs peroxisome proliferator-activated receptors (PPARs) α, β/δ, γ and the vitamin D receptor (VDR) and the identification of target genes of these receptors. The results emphasize the role of the most C-terminal helix of the LBD in the differences in co-activator (CoA) association and following transactivation. The stable fixing of this helix is possible in PPARs, as determined by their structural features, and enables ligand-independent CoA association. In contrast, VDR depends on its ligand to change to an active conformation. On the level of target genes, the family of insulin-like growth factor binding protein (IGFBP) genes was studied in detail in terms of responsiveness to PPARs and VDR. A meta-analysis of PPAR target genes addressed target gene features more broadly, including a detailed analysis of eight established PPAR targets. The binding site composition of the target genes of both receptors displayed an abundance of high affinity sites and a general trend towards enrichment of such sites in the regulatory areas. It was also evident that these NRs can function distally several thousands of base pairs from the transcription start site (TSS), both up- and downstream. Conservation trends assessed based on 38 known direct human PPAR targets provided insights into the turnover of binding sites and development of de novo binding sites that suggest that binding site composition is flexible. The detection of novel target genes was evaluated on human chromosome (chr) 19 genes and their mouse orthologs. In total six representatives from this panel were all shown to respond to PPAR ligands.

In conclusion, the studies presented in this doctoral thesis increased our understanding of the structure-function relationship of NR transactivation and typical features of binding site composition of target genes.

Universal Decimal Classification: 577.218, 577.214, 575.112, 575.113

National Library of Medicine Classification: QU 470, QU 475, QU 55, QU 34, QU 26.5

Medical subject headings: Genes; Gene Expression Regulation; Transcription Factors; Homeostasis; Nuclear Proteins; Receptors, Cytoplasmic and Nuclear; Peroxisome Proliferator-Activated Receptors; Receptors, Calcitriol; Insulin-Like Growth Factor Binding Proteins; Binding Sites; Computational Biology; Structure-Activity Relationship; Meta-Analysis
Acknowledgements

I would like to express my gratitude to all the people who motivated and helped me during the projects.

I acknowledge the Department of Biosciences and University of Kuopio for their role in hosting research facilities and providing high quality education. This research was also supported by funding obtained from the Academy of Finland, National Technology Agency of Finland (TEKES), the Finnish Cancer Organization and the European Union (Marie Curie RTN NucSys).

I would like to acknowledge my principal supervisor, Prof. Carsten Carlberg, first of all for including me in his research group already after my first year of studies, which motivated me to pursue a career in research. Moreover, the challenging scientific discussions and guidance over the years have greatly contributed to the creation of this thesis work.

I thank my other supervisors Prof. Jorma Palvimo and Sami Väisänen. Sami was always available for consultation on the laboratory methods and he has also had a major contribution in creating a good work atmosphere by hosting our group parties at his home.

I am grateful to the official reviewers of this thesis, Assistant Professor Sander Kersten and Prof. Alberto Muñoz, for their constructive comments and insights.

I thank my co-authors Ferdinand Molnár, Marjo Malinen, Katri Saavalainen, Tatjana Degenhardt and Oskari Uski for their contribution in the projects and their friendship, and Tom Dunlop and Karl-Heinz Herzig (A.I. Virtanen Institute, University of Kuopio) for their contribution in providing material from in vivo experiments. I would also like to acknowledge Tom for his critical reading of the manuscripts and this thesis text and for performing the language checking.

I am grateful also to the other present and former group members Anna, Tiina, Taru, Juha, Anssi, Antti, Aleksandra, Teemu, Harri L., Harri M., Lasse, Petra, Mikko, Manuel, Marc, Claire, Christian and Sabine for a friendly and motivating working atmosphere. Special thanks I owe to Christian who served as my supervisor in the lab when I started and with whom I also had many inspiring scientific discussions and Anna, Claire and Sabine for being close friends during these years.

I would like to thank the technical and administrative staff of the department, Marjatta Heikkinen, Liisa Alakuijala, and Taru Nylund. I am most grateful to our technician Maija Hiltunen for her contribution in providing and maintaining material for experimental work.

I acknowledge the support from my teaching colleagues Marjo, Jukka, Sanna, Petra and Markus, Prof Garry Wong, and the head of the Biochemistry Unit in the Department of Biosciences Annikka Linnala-Kankunen.

I am most grateful for the support and encouragement of my closest family, grandparents and my friends Tanja, Viivi, Aino and Jenni during my studies and this thesis work. I owe special gratitude to my fiancé Matti Heinäniemi for all of his contributions in supporting this thesis work and for being my personal stronghold, to whom I could always rely on.
**Abbreviations**

1α,25(OH)₂D₃  1α,25-dihydroxyvitamin D₃
Acyl-CoA  acetyl-Coenzyme A
ACOX1  acyl-CoA oxidase 1
ADRP  adipocyte differentiation related protein
AF-1  activation function 1
AF-2  activation function 2
ANF  atrial natriuretic factor
ANGPTL4  angiopoietin-like 4
APO  apolipoprotein
AR  androgen receptor
ARPO  acidic riboprotein 0
CAR  constitutive androstane receptor
CBP  cREB binding protein
CD36  CD36 antigen
CITCO  6-(4-chlorophenyl)imidazo[2,1-b][1,3]thiazole-5-carbaldehyde-O-3,4-dichlorobenzyl) oxime
ChIP  chromatin immunoprecipitation
Chr  chromosome
CoA  co-activator
CoR  co-repressor
COUP-TF  chicken ovalbumin upstream promoter transcription factor
COX  cyclooxygenase
CPT  carnitine palmitoyl transferase
CYP  cytochrome P450
DAX  DSS-AHC critical region on the X chromosome 1, gene 1
DBD  DNA-binding domain
DMSO  dimethylsulfoxide
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOTAP</td>
<td><em>N</em>-[1-(2,3-dioleoyloxy)propyl]-<em>N</em>,<em>N</em>,<em>N</em>-trimethylammonium methylsulphate</td>
</tr>
<tr>
<td>DR</td>
<td>direct repeat</td>
</tr>
<tr>
<td>ENCODE</td>
<td>Encyclopedia of DNA Elements</td>
</tr>
<tr>
<td>ER</td>
<td>everted repeat</td>
</tr>
<tr>
<td>ERα, ERβ</td>
<td>estrogen receptor α, β</td>
</tr>
<tr>
<td>ERR</td>
<td>estrogen-related receptor</td>
</tr>
<tr>
<td>FADS</td>
<td>fatty acid desaturase</td>
</tr>
<tr>
<td>FBS</td>
<td>fetal bovine serum</td>
</tr>
<tr>
<td>FXR</td>
<td>farnesoid X receptor</td>
</tr>
<tr>
<td>G0S2</td>
<td>G0/G1 switch gene</td>
</tr>
<tr>
<td>GCNF</td>
<td>germ cell nuclear factor</td>
</tr>
<tr>
<td>GK</td>
<td>glycerol kinase</td>
</tr>
<tr>
<td>GO</td>
<td>gene ontology</td>
</tr>
<tr>
<td>GST</td>
<td>glutathione S-transferase</td>
</tr>
<tr>
<td>GR</td>
<td>glucocorticoid receptor</td>
</tr>
<tr>
<td>GW501516</td>
<td>2-methyl-4-((4-methyl-2-(4-trifluoromethylphenyl)1,3thiazol-5-yl)-methylsulfanyl)phenoxy-acetic acid</td>
</tr>
<tr>
<td>GW7647</td>
<td>2-(4-(2-(1-cyclohexanebutylcyclohexylureido)ethyl)phenylthio)-2-methylpropionic acid</td>
</tr>
<tr>
<td>HAT</td>
<td>histone acetyl transferase</td>
</tr>
<tr>
<td>HDAC</td>
<td>histone deacetylase</td>
</tr>
<tr>
<td>HDL</td>
<td>high density lipoprotein</td>
</tr>
<tr>
<td>HepG2</td>
<td>human hepatoma derived cell line</td>
</tr>
<tr>
<td>HEK293</td>
<td>human embryonic kidney epithelial cell line</td>
</tr>
<tr>
<td>HNF4</td>
<td>hepatocyte nuclear factor-4</td>
</tr>
<tr>
<td>IGF</td>
<td>insulin-like growth factor</td>
</tr>
<tr>
<td>IGFBP</td>
<td>insulin-like growth factor binding protein</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>IL</td>
<td>interleukin</td>
</tr>
<tr>
<td>Insig</td>
<td>insulin signaling</td>
</tr>
<tr>
<td>IR</td>
<td>inverted repeat</td>
</tr>
<tr>
<td>LASS1</td>
<td>longevity assurance homolog 1</td>
</tr>
<tr>
<td>L783483</td>
<td>[3-chloro-4-[3-[7-propyl-3-(trifluoromethylenzo[d]isoxazol-6-yloxy)propylsulfonyl][phenyl]-acetic acid</td>
</tr>
<tr>
<td>LBD</td>
<td>ligand-binding domain</td>
</tr>
<tr>
<td>LBP</td>
<td>ligand-binding pocket</td>
</tr>
<tr>
<td>LDL</td>
<td>low density lipoprotein</td>
</tr>
<tr>
<td>LIPE</td>
<td>hormone sensitive lipase</td>
</tr>
<tr>
<td>LPL</td>
<td>lipoprotein lipase</td>
</tr>
<tr>
<td>LRH-1</td>
<td>liver receptor homolog 1</td>
</tr>
<tr>
<td>LRP</td>
<td>lipoprotein receptor-related protein</td>
</tr>
<tr>
<td>LXR</td>
<td>liver X receptor</td>
</tr>
<tr>
<td>MCF-7</td>
<td>human breast adenocarcinoma cell line</td>
</tr>
<tr>
<td>MR</td>
<td>mineralocorticoid receptor</td>
</tr>
<tr>
<td>NCoR</td>
<td>NR co-repressor</td>
</tr>
<tr>
<td>NGFIB</td>
<td>nerve growth factor inducible</td>
</tr>
<tr>
<td>NF-1</td>
<td>nuclear factor-1</td>
</tr>
<tr>
<td>NOR1</td>
<td>neuron-derived orphan receptor 1</td>
</tr>
<tr>
<td>NR</td>
<td>nuclear receptor</td>
</tr>
<tr>
<td>NURR1</td>
<td>nuclear receptor-related 1</td>
</tr>
<tr>
<td>p300</td>
<td>E1A-binding protein</td>
</tr>
<tr>
<td>PAK</td>
<td>p-21 activated kinase</td>
</tr>
<tr>
<td>PC-3</td>
<td>human prostate adenocarcinoma cell line</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>PEI</td>
<td>polyethylenimine</td>
</tr>
<tr>
<td>PEPCK</td>
<td>phosphoenolpyruvate carboxykinase</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Name</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>PG</td>
<td>prostaglandin</td>
</tr>
<tr>
<td>PGC-1α</td>
<td>peroxisome proliferators-activated receptor γ coactivator 1α</td>
</tr>
<tr>
<td>PIC</td>
<td>preinitiation complex</td>
</tr>
<tr>
<td>PNR</td>
<td>photoreceptor-specific nuclear receptor</td>
</tr>
<tr>
<td>PPAR</td>
<td>peroxisome proliferator-activated receptor</td>
</tr>
<tr>
<td>p-PolII</td>
<td>phosphorylated RNA polymerase II</td>
</tr>
<tr>
<td>PPRE</td>
<td>PPAR response element</td>
</tr>
<tr>
<td>PR</td>
<td>progesterone receptor</td>
</tr>
<tr>
<td>PXR</td>
<td>pregnane X receptor</td>
</tr>
<tr>
<td>RAC3</td>
<td>receptor-associated coactivator 3</td>
</tr>
<tr>
<td>RAR</td>
<td>retinoic acid receptor</td>
</tr>
<tr>
<td>Rb</td>
<td>retinoblastoma protein</td>
</tr>
<tr>
<td>RE</td>
<td>response element</td>
</tr>
<tr>
<td>ROR</td>
<td>retinoid orphan receptor</td>
</tr>
<tr>
<td>Rosiglitazone</td>
<td>5-[4-[(N)-methyl-N-(2-pyridyl)amino]ethoxy]benzyl] thiazolidine-2,4-dione</td>
</tr>
<tr>
<td>RXR</td>
<td>retinoid X receptor</td>
</tr>
<tr>
<td>RVR</td>
<td>rev-ERBA</td>
</tr>
<tr>
<td>SaOS-2</td>
<td>human bone osteosarcoma cell line</td>
</tr>
<tr>
<td>SF-1</td>
<td>steroidogenic factor 1</td>
</tr>
<tr>
<td>SHP</td>
<td>small heterodimer partner</td>
</tr>
<tr>
<td>SLC</td>
<td>solute carrier family</td>
</tr>
<tr>
<td>SOM</td>
<td>self-organizing map</td>
</tr>
<tr>
<td>SR-B1</td>
<td>scavenger receptor B1</td>
</tr>
<tr>
<td>SRC-1</td>
<td>steroid receptor coactivator 1</td>
</tr>
<tr>
<td>SSAT</td>
<td>spermidine/spermine N1-acetyltransferase</td>
</tr>
<tr>
<td>SULT</td>
<td>sulfotransferase</td>
</tr>
<tr>
<td>TF</td>
<td>transcription factor</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>TGF</td>
<td>tumor growth factor</td>
</tr>
<tr>
<td>Th</td>
<td>T helper cell</td>
</tr>
<tr>
<td>TIF2</td>
<td>transcription intermediary factor 2</td>
</tr>
<tr>
<td>TIMM</td>
<td>translocase of inner mitochondrial membrane</td>
</tr>
<tr>
<td>TLX</td>
<td>telomere (Drosophila)</td>
</tr>
<tr>
<td>TNFα</td>
<td>tumor necrosis factor α</td>
</tr>
<tr>
<td>TSS</td>
<td>transcription start site</td>
</tr>
<tr>
<td>TR</td>
<td>thyroid hormone receptor</td>
</tr>
<tr>
<td>TR2</td>
<td>nuclear hormone receptor TR2</td>
</tr>
<tr>
<td>TR4</td>
<td>nuclear hormone receptor TR4</td>
</tr>
<tr>
<td>UCP</td>
<td>uncoupling protein</td>
</tr>
<tr>
<td>UGT</td>
<td>UDP-glycosyltransferase</td>
</tr>
<tr>
<td>VDR</td>
<td>1α,25(OH)₂D₃ receptor</td>
</tr>
<tr>
<td>VDRE</td>
<td>1α,25(OH)₂D₃ response element</td>
</tr>
<tr>
<td>WM</td>
<td>weight matrix</td>
</tr>
<tr>
<td>WY14643</td>
<td>4-chloro-6-(2,3-dimethylphenylamino)pyrimidin-2-ylsulfanyl acetic acid</td>
</tr>
</tbody>
</table>
List of original publications

This dissertation is based on the following publications referred to in the text by their Roman numerals (I-IV):


*Equal contribution by both authors
## Table of contents

**Abstract** ......................................................................................................................... 1  
**Acknowledgements** ...................................................................................................... 3  
**Abbreviations** ............................................................................................................... 5  
**List of original publications** ...................................................................................... 11  
**1. Introduction** ............................................................................................................ 17  

**2. Review of Literature** ................................................................................................ 19  

- **2.1 Regulation of gene expression**  
  - 2.1.1 Development of the field of bioinformatics parallels the steps towards understanding the regulation of gene expression on a whole genome level ......................................................... 20  
  - 2.1.2 Human genome in comparison to genomes of other species ................................. 21  
  - 2.1.3 DNA architecture ......................................................................................................... 24  

- **2.2 Impact of bioinformatics on the understanding of the key elements of gene regulation: Regulatory proteins and DNA sequence elements**  
  - 2.2.1 Bioinformatic challenges at the protein level ............................................................ 26  
  - 2.2.2 Bioinformatics challenge on DNA level ..................................................................... 29  

- **2.3 NRs PPARα, γ, β/δ and VDR**  
  - 2.4 Structural aspects of NR function  
    - 2.4.1 Functional domains ................................................................................................. 34  
    - 2.4.2 Ligands ....................................................................................................................... 37  
    - 2.4.3 NR dimerization interfaces and DNA-binding ........................................................... 39  

- **2.5 Transactivation mediated by NRs**  
  - 2.5.1 Direct protein-protein interactions ........................................................................... 40  
  - 2.5.2 Regulation of transcription rates ............................................................................... 42  
  - 2.5.3 Dynamics of gene regulation by NRs ......................................................................... 42  

- **2.6 The role of PPARs and VDR on a genome wide scale**  
  - 2.6.1 Characterized physiology and known target genes of the VDR ............................ 43  
  - 2.6.2 Characterized physiology and target genes of PPARs ........................................... 47  
  - 2.6.2.1 PPARα ..................................................................................................................... 47
2.6.2.2 PPARγ ............................................................................................................................... 48
2.6.2.3 PPARβ/δ ........................................................................................................................... 50
2.6.3 Overlap in VDR and PPAR signaling.................................................................................. 52
2.7 Methods for the in silico prediction of NR binding sites and target genes 52

3. Aims of the Study....................................................................................................................55

4. Materials and methods........................................................................................................57

4.1 Materials ..........................................................................................................................57

4.1.1 Ligands ................................................................................................................................ 57

4.1.2 Cell lines ............................................................................................................................ 57

4.1.3 DNA constructs ................................................................................................................... 58

4.2 Methods ..................................................................................................................................59

4.2.1 In vitro methods ...................................................................................................................59

4.2.1.1 Transfection and luciferase reporter gene assay ............................................................... 59

4.2.1.2 In vitro transcription and translation ................................................................................. 60

4.2.1.3 Bacterial overexpression of proteins .................................................................................. 60

4.2.1.4 Gelshift and supershift assays.......................................................................................... 60

4.2.2 Ex vivo methods ..................................................................................................................61

4.2.2.1 RNA extraction, cDNA synthesis and real-time PCR ...................................................... 61

4.2.2.2 Chromatin immunoprecipitation assay ............................................................................. 62

4.2.2.3 PCR of chromatin templates ............................................................................................. 64

4.2.3 In vivo methods ...................................................................................................................64

4.2.4 In silico methods ................................................................................................................65

4.2.4.1 Structural analysis and visualizations of the protein structures ....................................... 65

4.2.4.2 In silico screening of putative PPREs .............................................................................. 65

4.2.4.3 Calculation of predicted binding strength ........................................................................ 65

4.2.4.4 Clustering using self-organizing maps (SOMs) ............................................................... 66

4.2.4.5 Sammon’s mapping ......................................................................................................... 66

5. Results .....................................................................................................................................69

5.1 Ligand-independent association of human PPARs with CoAs 69

5.2 Structural determinants of H12 stabilization in PPARs 70
5.3 Identification of direct VDR and PPAR target genes within the *IGFBP* gene family

5.3.1 Regulation of *IGFBPs* by VDR .............................................................. 72
5.3.2 Regulation of the *IGFBPs* by PPARs .................................................... 74

5.4 Meta-analysis of the binding site composition of human PPAR target genes 76

6. **Discussion** ...................................................................................................... 83

6.1 Ligand-independent CoA association of the PPARs 83
6.2 Structural determinants of H12 stabilization in PPARs 84
6.3 Identification of the direct VDR and PPAR target genes of the *IGFBP* gene family 85
   6.3.1 Regulation of the *IGFBP* gene family by VDR .................................... 86
   6.3.2 Regulation of the *IGFBP* gene family by PPARs ............................... 87
6.4 Meta-analysis of the binding site composition of human PPAR target genes 90

7. **Summary and conclusions** ............................................................................. 93

8. **Future aspects** .................................................................................................. 95

References .............................................................................................................. 97

**Appendix: Original publications**
1. Introduction

Biological systems show tremendous complexity at the level of structural organization and capacity to modulate their behavior in response to environmental changes. To achieve this, molecules with specific properties have evolved to store and transmit information, as well as carry out specialized functions (reactions, modifications, interactions). These molecules put together, determine what processes the system, such as a cell, a tissue, or an organ can carry out. At the same time, the lack of individual molecules or the change in concentrations of molecules in relation to each other can fundamentally alter the functioning of the system. Therefore, homeostasis is an important requisite for stable biological systems.

This stability requirement extends also to information storage and retrieval. In living systems information is stored in units called genes, that together can be used for the synthesis of about one order of magnitude more of the functional molecules of cells, the proteins they encode. As an example, humans have over 20,000 genes. Physically, genes exist as nucleic acids, most often as DNA molecules, and the information can be accessed via a process called transcription. Complexity has also evolved at the transcriptional level. The proteins responsible for the regulation and accessing of genetic information have also evolved to function together, in regulatory modules, or multi-protein complexes that possess many different activities. Their combined function is needed in response to environmental challenges. Therefore, homeostasis in humans involves and is maintained, via an integrated transcriptional control of multiple genes.

Recent research has brought vast advances in the characterization of whole sets of genes and proteins in different organisms. Luckily, in parallel advanced computational methods have been developed to access and analyze this data. Bioinformatics is a field of research that has developed the tools for comparing large molecular datasets, both DNA and protein, databases for storage of this data in a meaningful way and adapted statistical methods to analyze complex datasets and their biological relationships. The research in this field is tightly linked with the biological application of the methods, aiming to improve the computational models to suit the specific biological questions.

The functional properties of proteins are connected to their structural elements. The discovery of these relationships is the aim of structural informatics, where computational methods for visualization of structures and modeling of interactions can lead to the discovery of molecular mechanisms underlying their functionality. The challenge lies in the dynamics of the structures. For example, binding of a molecule, such as an activating ligand, to a family of TFs, called NRs, induces conformational changes in the whole molecule leading to altered interaction surfaces. At the level of DNA, higher organisms show a tendency of increased proportion of non-coding DNA. This region contains many of the regulatory elements that guide the first phase of accessing the gene information, transcription. These switches often correspond to binding sites of TFs. Subsequently a major sub-
discipline of bioinformatics research has focused on the development of methods that can faithfully identify these sites. However, locating the functional regulatory elements has proven challenging in mammalian genomes, owing to the large search space of gigabytes of potential binding sequence compared to binding site lengths of 5 to 15 bp. This means that in any given genome a large number of fits to the binding site sequence will occur. Since the number of TFs is limited in any cell, this leads to the situation that not all sites are used. Therefore one challenge has been to determine, which sites are functionally used in the set determined by the whole genome.

Ligand-activated NRs are key TFs that translate both nutritional and hormonal signals into changes in gene expression. Understanding the processes that they regulate as well as how they act is important for building a general understanding of how our body responds to nutrition, stress and thyroid and steroid hormones. Specific understanding will help to address possible treatment and prevention of disease states that origin from disturbance of this homeostatic control. As an example, the nutritional challenges today are very different from those in earlier times when food was scarce. It is important to acknowledge that our daily lifestyle affects the homeostatic control and may over time lead to the development of diseases, such as heart disease and type II diabetes.

This thesis aims to extend the research on NR regulomes, the characterization of the set of regulated genes and the functional properties affecting the regulatory process, with the application of bioinformatics methods, such as in silico binding site screening and structural analysis.
2. Review of Literature

2.1 Regulation of gene expression

Genes are DNA sequences that contain the information needed to make the functional units of the cell, its proteins, and comprise the information, the hereditary units, passed on every generation. Since the description of the DNA double helix by Watson and Crick in the early 1950s (Watson and Crick, 1953), the whole complexity of genomic events, the ways how our cells access and use this information, has started to become revealed. This flow of information is depicted in Fig. 1. We now acknowledge that DNA does not exist as a separate molecule in the cell. Instead it is wrapped around a protein core forming chromatin and also interacts with regulatory proteins.

![Diagram](image)

**Fig. 1. Information flow in the cell.** The genetic information is stored in the nucleus of each cell in the form of chromosomes. The chromosomes are long DNA molecules that contain genes, the instructions to make the functional molecules of cells, their proteins. The DNA sequence determines the amino acid sequence of proteins that finally fold to their final 3D-shape. The structural elements of proteins enable them to act alone or in complexes to perform many cellular functions (adapted from U.S. Department of Energy Human Genome Program, [http://www.ornl.gov/hgsc](http://www.ornl.gov/hgsc)).
Each cell type is defined by the proteins it produces. This means that each cell has a way to decide and determine, which genes get transcribed from DNA to RNA, and subsequently translated from RNA to protein. In a changing environment, however, it is necessary that the whole set of transcriptional events, starting from the level of which genes should be used, is dynamic. Moreover, cellular responses to environmental stimuli need to be coordinated and accurate. In effect, changes in the proteins produced are needed, which is achieved at the level of gene transcriptional regulation.

2.1.1 Development of the field of bioinformatics parallels the steps towards understanding the regulation of gene expression on a whole genome level

Information about DNA and protein sequences and structure started to build up in the 1950s. Databanks were established for the collection of structure and sequence data and in 1977 the first complete sequence of an organism, the bacteriophage \( \phi \) X174, was published (Sanger et al., 1977). With the arrival of the sequencing technique, more sequences rapidly accumulated from related organisms. This soon created a need for both storage and sequence comparison methods, which inspired the development of a new field, bioinformatics. The first algorithm for sequence comparison, the Needleman-Wunsch algorithm, had already been published in 1970 (Needleman and Wunsch, 1970), but it took more than ten years before more algorithms started to be published: Smith and Waterman published their algorithm for sequence comparison in 1981 (Smith and Waterman, 1981) and the FASTA algorithm was published in 1985 (Lipman and Pearson, 1985). Large central database organizations were also created at that time. The European Molecular Biology Laboratory (EMBL) is housing since 1981 the EMBL Data Library. In 1988 the National Center for Biotechnology Information (NCBI) was created in USA, which is responsible for similar sequence data repositories, such as Genbank. The third large DNA database was started in 1986 in Japan, called DDBJ. Today there is large variety of databases containing nucleic acid sequence data organized around specific qualities. These include the NCBI-housed refereed sequence database Refseq, the repbase repository for repetitive sequences and other types of databases concerning gene-disease associations, such as OMIM. More importantly, the end of 1980s also witnessed the commencement of the Human Genome Initiative. The human genome sequencing project was formally started in 1990, coinciding with the development of perhaps the most popular sequence comparison algorithm used today, BLAST (Altschul et al., 1990). During the 1990s several lower organism genomes were published, including bacterial \( E. \) coli and the nematode \( C. \) elegans genomes (Blattner et al., 1997; Waterston and Sulston, 1995). The decade of genome sequencing culminated in the publishing of the draft sequences of the human genome in 2001 by the public sequencing consortium and the private company Celera (McPherson et al., 2001; Venter et al., 2001).

The sequencing alone was a huge accomplishment, but with the sequence data at hand, new challenges emerged with bioinformatics playing a profound part to support both data storage and analysis. As a discipline that builds upon biosciences, mathematics, information science and statistics, the impact of bioinformatics is best seen as the application of these fields in ways that provide
insights into the vast, diverse and complex biological data making it more understandable and in trying to discern unifying principles. As an example, a vast number of unknown genes and proteins had been mapped from the genome. Deciphering their functional roles was something that algorithms for detecting sequence similarities or identifying functional domains of proteins, could help with (Birney et al., 2001). Database structures allowing functional classifications at various levels (such as Gene Ontology (GO) database) were created to unify data under common models to help understand the complex biological pathways and processes (an overview of the different databases currently existing is given in (Galperin, 2007)). Simultaneously, high-throughput molecular biology methods, such as microarrays, developed for the measurement of entire set of transcripts from cells (Ramsay, 1998), motivated the development of advanced data-analytical methods (Gollub and Sherlock, 2006). Together, the availability of sequence data and genome wide transcriptional analysis have made it possible to move from describing the genomic molecular parts list into the analysis of its dynamics and regulation. After the completion of the human genome sequence, the next large genomics task was launched in 2003 by the National Human Genome Research Institute (USA). The Encyclopedia Of DNA Elements (ENCODE) project was started to discover the remaining functional sequence elements of the human genome (in addition to gene sequences) through a collaborative effort of computational and laboratory-based scientists (ENCODE Project Consortium, 2004).

2.1.2 Human genome in comparison to genomes of other species

The human genome, the final version published in 2003 (International Human Genome Sequencing Consortium, 2004), consists of approximately three billion base pairs (3.08 Gb) of DNA, out of which only 1.2% is coding for proteins (Table 1). A large fraction of the genome consist of repetitive sequences, interspersed elements already occupy 45% of the genome and may have a less significant functional role. The remaining sequence is used for regulation of gene expression and genome organization. In total, humans have approximately 20,000 to 25,000 genes (latest release of the European database Ensembl contains 22,205 genes). These can be classified according to their role to different biological processes. This labeling of genes by functional classification is the function of the GO database. The largest category of known genes consists of genes encoding proteins that regulate and maintain the DNA information, TFs and other DNA-binding proteins (Table 2). This is followed by the category signal transduction. This data emphasizes the importance placed in understanding the communication between environmental signals and gene transcription. Surprisingly, still three years after the completion of the human genome sequence, a full consensus around the complete set of coding genes does not exist. Only very few of the human genes are thought to be missing from the computational predictions, but the exact genomic structure is estimated to be correct for only 50% of the predicted genes, due to the large number of alternative splicing sites in the human genes (Guigo et al., 2006).
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes

Table 1. Overview of the human genome

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of the genome (Gb)</td>
<td>3.08</td>
</tr>
<tr>
<td>Haploid chromosome number</td>
<td>23</td>
</tr>
<tr>
<td>Fraction of coding sequence (%)</td>
<td>1.2</td>
</tr>
<tr>
<td>Number of genes</td>
<td>20,000-25,000</td>
</tr>
<tr>
<td>Average gene size (kB)</td>
<td>27</td>
</tr>
<tr>
<td>Most gene rich chromosome (genes/MB)</td>
<td>Chr 19 (23)</td>
</tr>
<tr>
<td>Least gene rich chromosome</td>
<td>Chr 13, Chr Y (5)</td>
</tr>
</tbody>
</table>

Table 2. GO-categories involved in gene regulation and signaling in humans

<table>
<thead>
<tr>
<th>GO-category</th>
<th>% of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleic acid binding (total)</td>
<td>13.5</td>
</tr>
<tr>
<td>Transcription factor</td>
<td>6.0</td>
</tr>
<tr>
<td>Nucleic acid enzyme</td>
<td>7.5</td>
</tr>
<tr>
<td>Signal transduction (total)</td>
<td>12.2</td>
</tr>
<tr>
<td>Signaling molecule</td>
<td>1.2</td>
</tr>
<tr>
<td>Receptor</td>
<td>5.0</td>
</tr>
<tr>
<td>Kinase</td>
<td>2.8</td>
</tr>
<tr>
<td>Regulatory molecule</td>
<td>3.2</td>
</tr>
</tbody>
</table>

The best-characterized eukaryotic genome to date still is the yeast genome (Goffeau et al., 1996). Compared to this simple eukaryote (genome size 12 MB), the genome in humans has expanded profoundly (3 GB). Additionally, the complexity of its regulation and maintenance has increased as well. Yeast has in average one regulatory protein per 20 genes, while humans have one for every ten genes (Levine and Tjian, 2003). To date, the regulation of gene expression in yeast has been successfully characterized in a relatively detailed manner with respect to participating molecules and regulatory sequences. It has also been possible to create lists of target genes for each regulatory molecule, termed regulomes. In comparison, only the initial steps towards the human regulome have been made. The ENCODE project has started to map the regulatory DNA of selected chromosomal locations of the human genome down to a base pair resolution. The pilot phase is focused on a selected 30 MB fragment of the genome, representing approximately 1% of the whole sequence. This is already almost three times larger than the whole yeast genome.

Soon after the human genome was finished, the mouse and rat draft genomes became available (Gibbs et al., 2004; Waterston et al., 2002), followed by a number of different species including our closest relatives (measured by sequence similarity) in the family of great apes (Chimpanzee Sequencing and Analyzing Consortium, 2005). The bioinformatics analysis utilizing these other mammalian and vertebrate genomes, comparative genomics, has added extra dimensionality to the understanding of how mammalian genomes evolved as well as how the human genome differs in
form and function from the others (Table 3). On the level of genes and proteins humans are very similar to other mammals. Changes in the genomes happen via a spectrum of events, ranging from the very large down to the single base pair level (the smaller changes will be discussed later). The largest changes occurred mainly via rearrangements of chromosomes. This is evident from the comparisons of gene order in the different mammalian species: typically, large segments of chromosomes retain gene order. Due to these large rearrangements, however, it can be seen that regions of conserved gene order have ended up on different chromosomes in today’s mammalian genomes. These corresponding segments in the other genome are described as being syntenic. In general, these rearrangements are rare events in genomes, one every ten million years, with the exception of the genomes of rodents, dog, cow and some of the New World apes that show very high rates of rearrangements relative to other mammals (Murphy et al., 2001).

Table 3. Characteristics of selected vertebrate genomes studied to date (adopted from (Murphy et al., 2001))

<table>
<thead>
<tr>
<th>Species</th>
<th>Common name</th>
<th>Haploid chr number</th>
<th>Genome size (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>Chimpanzee</td>
<td>24</td>
<td>3.1</td>
</tr>
<tr>
<td>Gorilla gorilla</td>
<td>Gorilla</td>
<td>24</td>
<td>NA</td>
</tr>
<tr>
<td>Pongo pygmaeus</td>
<td>Orangutan</td>
<td>24</td>
<td>NA</td>
</tr>
<tr>
<td>Cercopithecus</td>
<td>African green monkey</td>
<td>30</td>
<td>NA</td>
</tr>
<tr>
<td>aethiops</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cebus capucinus</td>
<td>Capuchin monkey</td>
<td>27</td>
<td>NA</td>
</tr>
<tr>
<td>Papio hamadryas</td>
<td>Baboon</td>
<td>21</td>
<td>NA</td>
</tr>
<tr>
<td>Macaca mulata</td>
<td>Rhesus monkey</td>
<td>21</td>
<td>3.1</td>
</tr>
<tr>
<td>Lagomorpha</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oryctolagus cuniculus</td>
<td>Rabbit</td>
<td>22</td>
<td>3.5</td>
</tr>
<tr>
<td>Rodentia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mus musculus</td>
<td>Mouse</td>
<td>20</td>
<td>2.5</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>Rat</td>
<td>21</td>
<td>2.8</td>
</tr>
<tr>
<td>Carnivora</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>Dog</td>
<td>38</td>
<td>2.4</td>
</tr>
<tr>
<td>Felis catus</td>
<td>Cat</td>
<td>19</td>
<td>3.0</td>
</tr>
<tr>
<td>Cetartiodactyla</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bos Taurus</td>
<td>Cow</td>
<td>30</td>
<td>3.0</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>Pig</td>
<td>19</td>
<td>2.8</td>
</tr>
<tr>
<td>Galliformes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>Chicken</td>
<td>37</td>
<td>1.3</td>
</tr>
<tr>
<td>Atheriniformes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Danio rerio</td>
<td>Zebrafish</td>
<td>25</td>
<td>1.7</td>
</tr>
</tbody>
</table>
Comparisons of the human and the chimpanzee genome have allowed the assessment of key differences between humans and their closest relatives. The genome-wide divergence of human and chimpanzee is 1.06%, so for the most part the comparisons between these species revealed similarities. Compared to other mammalian genomes, together the hominid genomes show a rapid evolution in a subset of genes involved in immunity, such as those responsible in antimicrobial activity, olfactory genes and genes functioning in reproduction. In contrast, processes including intracellular signaling, metabolism and especially genes expressed in the brain show low rates of evolution. However, compared to mice, accelerated evolution in hominids involves also genes with brain-related functions. Interestingly, the comparative analysis of human and chimpanzee also revealed instances, where the human disease-causing allele is the wild type allele in chimpanzee. For example, proline is the wild type amino acid in position 12 of the NR PPARγ, in humans this variation has been linked with increased risk of type II diabetes (Chimpanzee Sequencing and Analyzing Consortium, 2005).

The mouse genome, which diverged from the human genome 75 million years ago, is most often used in sequence comparison to detect functional conservation, especially conservation of regulatory sequence elements. Relatively large proportion of genes can be matched between these species: approximately 80% of mouse genes have an ortholog, a gene encoding a protein with similar structure and function, in the human genome (Bulyk, 2003). The main motivation of utilizing the mouse genome for sequence comparisons instead of closest relatives to humans, such as the chimpanzee genome, is the sufficient evolutionary distance, in other words the sequence had enough evolutionary time for random mutations to separate well-conserved functional regions outside coding area. More recently, comparisons between related species has been suggested as an alternative approach to study lineage-specific conservation. This phylogenetic footprinting or shadowing aims to identify essential elements by identifying differences in closely related species. A collection of related species, such as several primate genomes, are compared together, which enables localization of collective variable regions and conserved regions. The utility of this approach was studied by Boffelli and colleagues with the example of intron-exon boundaries of several genes and to detect the regulatory regions of a human-specific gene, apolipoprotein (a) (Boffelli et al., 2003).

It is interesting to try and relate the sequence differences to the phenotypic differences observed between species. However, the comparative analysis has shown that no direct relationships exist between sequence diversity and phenotypic diversity. Humans differ phenotypically much more from chimpanzee than two closely related mouse species, however, dogs show very little sequence divergence but have highly different phenotypes. It is possible that a number of small changes correspond the variation observed. Never the less, it is also intriguing to suggest that part of the differences in the end come from transcriptional regulation.

2.1.3 DNA architecture

DNA exists in the cells wrapped around a core of histone proteins (a core octamer formed from two copies of histones H2A, H2B, 3 and 4 and histone H1 covering the linker bases) with approximately
In this form, DNA can be condensed to the very extreme form of metaphase chromosomes during cell division or exist in various intermediate states that by physical means can restrict or allow the accessibility to certain regions of the genome. Between cell divisions two main forms can be distinguished: accessible open chromatin, or euchromatin, and inaccessible densely packed heterochromatin. The functional properties of histones can be regulated to affect the local chromatin status. A common way to modify the function of proteins is via covalent modifications. This also applies to the histone core proteins of the nucleosome, for which several modifications have been described (Grant, 2001). The acetylation of both histones H3 and H4 are well-known labels of active chromatin, whereas methylation of histone H3 at the lysine residue at position 9 is one example of a modification that leads to inactivation and favors heterochromatin formation. Both acetylation and methylation act by either adding hydrophobicity to the histone tail or by smothering it’s positive charge thus alleviating the strong interaction between positive charge on lysines and the negatively charged DNA wrapped around it. Alleviating the interaction favors access of regulatory proteins and subsequent recruitment of other complexes that can, for example, shift nucleosome positions to permit further access to regulatory DNA elements. In contrast, the methylation of histone H3 lysine 9 creates a binding site for heterochromatin forming proteins leading to silencing of transcription.

In total, human genetic information is spread over 23 pairs of linear DNA molecules (chromosomes) that vary in length from 47 to 247 MB, all contained in the nucleus of the cell. It has been proposed that the spatial location in the nucleus reflects also chromatin status, with inactive chromatin mainly occupying nuclear periphery (Swedlow and Lamond, 2001). In support of this hypothesis, the inactive X chromosome in females, also known as the Barr body, exists in heterochromatin form in the nuclear periphery. It is also known that the gene density varies greatly between chromosomes. A study comparing the gene-poor chr 18 and the gene-rich chr 19 showed that chr 18 was located near nuclear periphery, while after inhibition of transcription the preferential localization disappeared (Croft et al., 1999). The localization may be explained by recruitment of heterochromatin-specific proteins to inactive sites, which direct the localization pattern.

There is evidence that chromatin is organized in loops that get attached to the nuclear matrix (approximately 100 kB in length) and these chromatin units are the functional domains, within which regulatory elements can come into contact with the genes in that domain by DNA looping. Simonis and colleagues (Simonis et al., 2006) demonstrated recently that active and inactive genes are connected via long-range interactions. Preferentially, active genes are interacting together, similarly inactive ones cluster to the same region. The locations of individual genes are also likely determined by the types of proteins binding to their regulatory sequences, the binding of TFs directing the genes to active regions. Furthermore, Spilianakis and co-workers showed evidence of an interchromosomal regulation, where even DNA elements existing on different chromosomes could regulate genes that were physically attached to the same nuclear region (Spilianakis et al., 2005).
2.2. Impact of bioinformatics on the understanding of the key elements of gene regulation: Regulatory proteins and DNA sequence elements

The regulation of gene expression involves a large set of nuclear proteins that includes the chromatin modifying proteins, together with TFs, co-regulators and the basal transcriptional machinery (Levine and Tjian, 2003). A central role in the process, the way how environmental stimuli, or other signals from the cell are translated into changes in gene expression, is mediated by the activation and binding of DNA sequence selective TFs to the regulatory regions of a gene. This recognition event often leads to the recruitment of multi-protein complexes to these sites. Included with these macromolecular assemblies are co-regulator proteins that in turn recruit other multi-factor complexes that carry out tasks, such as chromatin re-modeling, or connecting the TF to the basal transcriptional machinery. The basal transcriptional machinery consists of several proteins that bind next to the TSS of genes and enable RNA polymerase II-mediated gene transcription to occur. In summary, regulation of gene expression involves a series of intermolecular interactions that starts from a specific DNA-protein contact at each TF binding site and extends to protein-protein interactions enabled by structural modifications. This finally leads to a change in the number of departures of the RNA polymerase from the TSS of a gene (i.e. the transcription rate). Importantly, this change is simultaneously taking place for the entire set of genes that can be recognized by the TF in a given tissue.

The task to define the set of genes regulated by each TF, its regulome, involves the understanding and characterization of the following aspects of the process: Firstly, the initiating DNA-protein interaction involves a recognition of a specific DNA sequence, which can be defined for each TF. Secondly, the following intermolecular interactions encompass a large selection of regulatory proteins, where specificity and preferential interactions may dictate the overall dynamics of the process, affecting the time scale, during which a change in chromatin status and the assembly of transcriptional machinery are observed for a given TF. Finally, the overall result is the change in transcriptional rate of each target gene that will lead to a change in the respective mRNA levels. As a consequence, two bioinformatics challenges can be identified in the characterization of regulomes 1) understanding at the protein level the functional properties of participating molecules by modeling as well as visualizing intermolecular interactions and 2) identifying on a genomic scale the transcriptional targets by binding site detection algorithms and microarray methods.

2.2.1 Bioinformatic challenges at the protein level

Multiple databases have been created that collect protein sequences, perform classification and similarity searches with final documentation of the findings. In addition, protein crystal structures have been systematically stored into the Protein Data Bank (PDB). As more and more structures become available, structural comparisons and prediction of protein structures by modeling has inspired the development of dedicated software in the field of bioinformatics. This has resulted in advances in the understanding of the relationship between protein structures and their functional properties. The process of protein crystallization requires time and effort and is not readily adaptable.
to the high-throughput scales of other techniques, such as microarrays. Subsequently, the prediction of functional properties by modeling the structure using structural neighbors can be an attractive alternative (Tramontano, 2006).

Computer-aided visualization of the determined protein structures, or models, permits the identification of functional parts of the protein, such as different domains, catalytic parts, modifiable side chains and protein-protein interaction surfaces. Many proteins bind small molecules (such as substrates, ligands and prosthetic groups). Examination of their binding mode is of special interest in pharmaceutical industry for the development of drug molecules that may serve as activators or inhibitors of these protein structures. It can be also of use in the understanding of the activation mechanisms perpetrated by TFs, in particular the subgroup of ligand-activated TFs, such as the NRs. Ligand-binding induces structural changes in these molecules that are evident from the examination of bound/unbound crystal structures. These changes translate to a modification of protein-protein interaction surfaces, which in turn explain transactivation mechanisms of different family members (Robinson-Rechavi et al., 2003). Activation by ligand is only one way of affecting the structural properties of a TF. Identification of covalent modification, such as phosphorylation, acetylation or the more recently characterized sumoylation of TFs, leads to a better understanding of the functional consequences that these modifications will have (for example, those described in the NF-κB TF signaling pathway (Perkins, 2006)). Phosphorylation is a very widespread modification, especially of those TFs that are activated by signaling cascades initiated at the cell surface by various ligand-specific receptors, such as cytokine signaling via the Jak/STAT pathway (reviewed in (Imada and Leonard, 2000)).

The nature and the selectivity of the protein-DNA interaction are of special interest in understanding gene regulation. The PDB now contains the structures of several DNA-binding protein classes. These can be divided into eight different structural/functional groups (Luscombe et al., 2000): the helix-turn-helix, zinc coordinating, zipper-type, other α-helix, β-sheet, β-hairpin/ribbon, enzyme-type DNA-binding proteins and others (Fig. 2). Homeodomain TFs are an example of the group of helix-turn-helix proteins. Zipper-type proteins include TFs, such as Fos, Jun and Max. However, in eukaryotes the single largest group of TFs is the zinc co-ordinating group. These proteins utilize a metal ion in their DNA-binding domain (DBD). The role of the metal ion is believed to create structural stability in domains that are not sufficiently large for a stable hydrophobic core. The metal ion itself is held in place by interactions with discretely positioned amino acid residues, either cysteines or histidines. TFs, such as NRs, the zinc-finger proteins and the tumor suppressor protein p53 belong to this group. Additionally, two groups of proteins utilize β-sheets in DNA-binding. The β-sheet proteins include, for example, the TATA-box-binding protein (a component of the basal transcriptional machinery). Finally, the group named as ‘other’ includes the STAT family of TFs and Rel TFs. The details of the DNA-binding mode of the different classes of TFs are extensively reviewed in the literature (Luscombe et al., 2000). However, given the structural diversity of TFs, no simple rule exists on how the amino acid sequence relates to the DNA sequence it binds. Family-specific rules for DNA base recognition can be generated, though, by examining the structures.
Moreover, via modeling of the binding of different DNA sequences to those TFs that have a solved crystal structure (or closely related crystal structure) available, relatively accurate predictions of binding strength can be made (Morozov et al., 2005).

**Fig. 2. Families of DNA binding proteins.** The structural elements to recognize DNA have evolved diversity. Representatives from each class of DNA binding proteins are shown with the reference PDB code in brackets. Adopted from (Luscombe et al., 2000).
The group other α-helix contains the histone proteins alongside with some TFs. Recently, the DNA-interaction of histone molecules was modeled in order to predict histone locations on DNA sequences of different base pair composition (Segal et al., 2006). Such predictions could be useful in predicting the physical accessibility of individual DNA regulatory elements \textit{in vivo}. As a result this model successfully predicted histone-poor regions (such as those with high CG-content). Moreover, a tolerable overall fit was obtained between the model and experimental verification over longer sequences around transcribed genes.

\subsection*{2.2.2 Bioinformatics challenge on DNA level}

By experimental means, or by structural modeling techniques, it is possible to address the question, what kind of DNA sequences are recognized by each TF. The recognition sequences vary in length between 5 and 15 bp and due to the sheer size of the mammalian genome in effect have a multitude of occurrences within the 3 billion base pair sequence. Methods, such as gelshift (or electrophoretic mobility shift assay), SELEX, or more recently developed array based assays can be used to screen/compare binding of different DNA elements. It is important to note that \textit{in vitro} determined binding does not necessarily represent binding strength \textit{in vivo}, where the presence of other binding factors may alter stability. Secondary interactions between TFs can enable even weak affinity sites to be functional in the proper context and the exact contribution of, such effects is difficult to address experimentally (Bulyk, 2003; Hallikas et al., 2006).

Since the completion of the sequencing of the human genome, it has become possible to search for defined binding motifs from entire genomes. For this purpose, bioinformatics research has developed different ways, in which to represent the binding site motif in search algorithms that will predict possible binding sites. The most often used approach is the weight matrix (WM) approach, where known, experimentally defined binding sites collected from publications are aligned and used to form a matrix that represents the frequencies of the four DNA bases (A, G, C, or T) at each location in the binding site (Stormo, 2000). A threshold is then determined for the matrix that will identify putative binding sites most similar to those already known. The ease of the use of matrix-based methods and their ready availability from databases has made them popular. However, the results themselves are not in direct relationship to binding strength and the performance (and quality) of the matrix is also dependent on the availability of sufficient number of known binding sites.

Recently, alternative methods that exploit experimental binding data sets were developed (Hallikas et al., 2006). These scoring matrices perform well in the detection of near consensus RE variants and also approximate to a certain extent their relative binding strengths. A study by Morozov and colleagues compared these methods to the structural informatics approach, a computer simulation of the binding of different DNA sequences to the TF molecule (possible only where a crystal structure of the TF DNA-binding domain exists) (Morozov et al., 2005). In this work, they showed that as the
number of variations increases, the combinatorial effects, not predictable when considering single nucleotide positions, also start to emerge.

The availability of a number of complete genomes has motivated the development of binding site prediction tools that exploit sequence alignments between several species to detect potential functional regions in the genome of interest, with the hypothesis that evolutionary pressure maintains elements and that this implies functionality. Mammalian regulatory regions often involve a set of TF binding sites clustered together, called enhancer regions. It is thought that several such TF modules can exist per gene and they can be located distally relative to the TSS. The functionality of the module is often dependent on the presence of several TFs and therefore a larger than 5 to 15 bp window conserved between species is searched for, assuming that conservation is evidence for functional importance. Visel and co-workers reported a database search tool, called EnhancerBrowser, for such conserved regions between different genomes using whole genome alignment methods (Visel et al., 2007). Moreover, it has been proposed that inside such modules cooperative binding is also possible, leading to the stabilization of TFs to weak binding sites. This can be a desired feature, rather than creating a platform for any of the participating TFs to bind (an OR-type switch), the requirement of stabilization creates an exclusive AND-type switch, where activation occurs only in the presence of all factors.

For the conservation of functionality it may not be necessary that the exact order of TF binding sites is conserved (which is assumed in global alignment approaches). Hallikas and colleagues reported conserved enhancer regions defined by the preservation of the same TF binding site composition, irrespective of the surrounding sequence in total, in their database of human to mouse sequence alignments that was created by the software Enhancer Element Locator (EEL) (Hallikas et al., 2006). The stabilization hypothesis is a key element in their enhancer locator, a requirement that restricts the distribution of TF binding sites forming the module. The concept of binding site clustering had been explored before in Drosophila (Ludwig et al., 2000). Based on the comparison of two closely related fly species, D. melanogaster and D. pseudoobscura, Berman et al. proposed a very similar enhancer identification software relying on TF clustering (Berman et al., 2004). Based on the fly studies conservation of the similar TF modules was not apparent, rather locations, in addition to patterns of TF sites, were distinct. The clustering approach alone, however, leads to a higher number of false positive sites, when applied to more complex mammalian genomes (Hallikas et al., 2006).

A study by Ludwig and co-workers compared enhancers of distal and closer relative of D. melanogaster (Ludwig et al., 2005). Substituting the enhancer from a distant relative species D. pseudoobscura, resulted in an identical pattern production driven by the expression of the gene eve. This is surprising since, the enhancer is completely conserved in only three of the 18 binding sites and two are lacking completely. On the other hand, the enhancer from the closer relative D. erecta, failed to preserve eve gene expression. This suggests that in contrast to amino acid sequence of proteins, enhancers show a more flexible architecture that allows modification, including turnover of binding sites. Different rules of molecular evolution thus apply to regulatory sequences. Furthermore,
the differences between *D. melanogaster* and *D. erecta* enhancers may reflect differences in TF concentrations that activate the enhancer.

*De novo* binding sites can readily evolve (Stone and Wray, 2001). This rapid turnover of binding sites was demonstrated with a computer simulation of factors controlling evolution of binding sites. The waiting time, measured within a neutral evolution model involving random mutations, was shown highly dependent on sequence length (exponential scaling) and more moderately to be dependent on the length of regulatory area (linear scaling). With a population size of one million, a novel hexameric binding site was shown to evolve in every 2,250 generations, which in terms of evolution is microscale. The preservation of this new binding site, its fixation, was also estimated. Taking the fruit fly as an example, for a generation time of five weeks, it would only take 75 years to establish a new binding site. Given that species like *D. melanogaster* and *D. pseudoobscura* diverged 40 to 60 million years ago, the previous findings of Ludwig and colleagues, that the enhancers, while retaining functionality, have diverged considerably, is not that surprising. Even the simultaneous appearance of two binding sites within 200 bp region was calculated to take only 55,000 years, which is a small time interval in terms of evolution.

Instead of searching for functional binding sites or modules, the outcome of regulation is another starting point to define regulomes. This requires determining mRNA levels from cells under two conditions, comparing the activation of a TF to the normal state, and the subsequent analysis of the complex dataset. The dynamic nature of the regulation means that in order to capture all events, a time series experiment is needed. This typically results in different types of response according to which the responding genes can be clustered using sophisticated analysis tools. The clusters themselves can then serve as input for algorithms to discover similarities and differences in the regulatory elements (Huber and Bulyk, 2006). Unfortunately, this approach alone does not distinguish indirect target genes that depend on the activation of some of the direct targets but themselves lack binding sites for the TF in question. As a solution, recently developed microarray-based immunoprecipitation techniques enable the experimental detection of bound TFs to the non-repetitive sequence component of the genome (Horak and Snyder, 2002). However, in multi-cellular organisms, such as humans, no single cell type in the adult is expressing the whole set of genes. As a consequence, the full regulome for a TF needs to be assembled from an overlap of a set of different tissues. Holstege and Clevers compared these bioinformatics and experimental approaches (Holstege and Clevers, 2006). Clearly, the understanding of regulation on a genome-wide scale has advanced by the development of more efficient methods in both fields, yet further challenges need to be overcome before a complete description of regulatory regions can be compiled.

The modular enhancer type activation model suggests that regulation of transcription is often dependent on the presence of a set of TFs co-operatively playing a role in the regulation. In effect, the process to build a regulome for a certain TF, in the end, translates into a process of describing interconnected regulomes. This requires the collection of extensive datasets that describe binding site preferences, location preferences and expression profiles for thousands of TFs. Taken together, the integration of various types of evidence, both experimental and computational, will be required in the
characterization of regulomes. This continues to motivate the tight interaction between computational biology and wet-lab molecular biology.

2.3 NRs PPAR\(\alpha\), \(\gamma\), \(\beta/\delta\) and VDR

NRs form a TF family of 48 members in humans. Some of the family members are responsive to nanomolar range concentration of activating ligands, another group senses their ligands at micromolar range, whereas for the remaining receptors, a natural ligand has not been characterized (Chawla et al., 2001b). The ligand-centered view (Table 4) suggests that ligand responsive NRs bind small hydrophobic compounds ranging from bulky xenobiotic compounds via different steroid structures to more simple fatty acid derivatives. Being lipophilic, the ligands can readily enter cells through membranes and reach their receptor inside the cell. Many NRs are located inside the nucleus and only some NRs reside in the cytoplasm.

Alternatively, evolutionary relationships (at the primary amino acid sequence level) reveal a different picture of the family (Fig. 3) (Bertrand et al., 2004), which is more focused on differences that translate into the structural and functional variations. One example is different dimerization modes, since a large subgroup of NRs forms dimeric complexes with RXRs (indicated with a star). Recently, the comparison of the expression levels and distribution of all the NRs in mice provided yet another way of classification that may better reflect shared function and physiology of the family members (Fig. 3) (Bookout et al., 2006). This study demonstrated that groups of NRs play central roles in the long-term regulation of cellular control processes, such as inflammation, cell growth and differentiation, nutrient storage and use, development and reproduction.

The human VDR was cloned in 1988 (Baker et al., 1988). This NR is classically associated with the regulation of calcium homeostasis and bone formation but is also known to be involved in a number of other physiological processes, especially its role in cell growth and inflammation has been extensively studied (Hansen et al., 2001; Levine and Tjian, 2003). In the early 1990s, NRs sensing fatty acids, called PPARs, were cloned from different species. The first member, PPAR\(\alpha\), was found responsive to peroxisome proliferators, hence the name and indeed the activation of this receptors leads to an increase in the number of peroxisomes in the liver of rodents (Issemann and Green, 1990). In humans, this effect is lacking and beneficial roles have been shown in several studies examining epidemiologic data from clinical use of PPAR\(\alpha\) ligands in the treatment of disease (Tenkanen et al., 1995; Robins et al., 2003). Peroxisomes play a role in the oxidation of fatty acids that also occurs in the mitochondria. Several genes in these oxidative pathways and in pathways of fat transport, distribution and storage have later been characterized as PPAR target genes. Shortly after the first member was discovered, the isotypes, PPAR\(\beta/\delta\) and PPAR\(\gamma\), were cloned (Dreyer et al., 1992; Kliewer et al., 1994). Both VDR and PPARs have attracted the attention of pharmacological research, VDR for its role for treatment of hyper-proliferative diseases and PPARs as key regulators of energy metabolism (especially fat but also influencing glucose levels). Several drugs, such as thiazolidinediones (TZDs), that are currently used to treat type II diabetes, or fibrates that are used for
the treatment of dyslipidemia are in fact activators for these receptors. As drug targets, these receptors are of interest for structure/function characterization and the widespread physiological roles make them good targets for regulomics research.

Table 4. Classification of NRs based on ligand binding

<table>
<thead>
<tr>
<th>Subfamily</th>
<th>Name</th>
<th>Nomenclature</th>
<th>Ligand*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AR</td>
<td>NR3C4</td>
<td>dihydrotestosterone</td>
</tr>
<tr>
<td></td>
<td>ERα, β</td>
<td>NR3A1, A2</td>
<td>17β-estradiol</td>
</tr>
<tr>
<td>Receptors with high</td>
<td>GR</td>
<td>NR3C1</td>
<td>cortisol</td>
</tr>
<tr>
<td>affinity for ligands</td>
<td>MR</td>
<td>NR3C2</td>
<td>aldosterone</td>
</tr>
<tr>
<td></td>
<td>PR</td>
<td>NR3C3</td>
<td>progesterone</td>
</tr>
<tr>
<td></td>
<td>RARα, β, γ</td>
<td>NR1B1, B2, B3</td>
<td>all-trans retinoic acid</td>
</tr>
<tr>
<td></td>
<td>TRα, β</td>
<td>NR1A1, A2</td>
<td>3,5,3’-triiodothyronine</td>
</tr>
<tr>
<td></td>
<td>VDR</td>
<td>NR1H1</td>
<td>1α,25-dihydroxyvitamin D3</td>
</tr>
<tr>
<td>Sensors with low affinity for ligands</td>
<td>CAR</td>
<td>NR1H3</td>
<td>androstanol, xenobiotics</td>
</tr>
<tr>
<td></td>
<td>ERα, β, γ</td>
<td>NR3B1, B2,B3</td>
<td>anti-estrogens</td>
</tr>
<tr>
<td></td>
<td>FXR</td>
<td>NR1H4</td>
<td>bile acids</td>
</tr>
<tr>
<td></td>
<td>HNF-4α, γ</td>
<td>NR2A1, A2</td>
<td>fatty acids</td>
</tr>
<tr>
<td></td>
<td>LRH-1</td>
<td>NR5A2</td>
<td>phospholipids</td>
</tr>
<tr>
<td></td>
<td>LXRα, β</td>
<td>NR1H2, H3</td>
<td>oxysterols</td>
</tr>
<tr>
<td></td>
<td>PPARα, γ, β/δ</td>
<td>NR1C1, C2, C3</td>
<td>fatty acids and their derivatives</td>
</tr>
<tr>
<td></td>
<td>PXR</td>
<td>NR1H2</td>
<td>pregnanediol, xenobiotics</td>
</tr>
<tr>
<td></td>
<td>RORα, β, γ</td>
<td>NR1F1, F2, F3</td>
<td>fatty acids, cholesterol, retinoids</td>
</tr>
<tr>
<td></td>
<td>RXRα, β, γ</td>
<td>NR2B1, B2, B3</td>
<td>retinoids, fatty acids</td>
</tr>
<tr>
<td></td>
<td>SF-1</td>
<td>NR5A1</td>
<td>phospholipids</td>
</tr>
<tr>
<td>Orphans</td>
<td>COUP-TFα, β, γ</td>
<td>NR2F1, F2, F6</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>DAX</td>
<td>NR0B1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCNF</td>
<td>NR6A1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NGFIB, NURR1, NOR1</td>
<td>NR4A1, A2, NR4A3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PNR</td>
<td>NR2E3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RVRα, β</td>
<td>NR1D1, D2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SHP</td>
<td>NR0B2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TLX</td>
<td>NR2E1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TR2, TR4</td>
<td>NR2C1, C2</td>
<td></td>
</tr>
</tbody>
</table>

*Classification based on synthetic or natural compounds co-crystallized with NRs, physiological effect for all sensor NRs not established
2.4 Structural aspects of NR function

2.4.1 Functional domains

The functional properties of NRs depend on their structural elements. The majority of the NRs contain the following domains that have similar functions in each receptor (Mangelsdorf et al., 1995). The N-terminus of most NRs hosts the activation function 1 (AF-1) domain that has been characterized as one possible contact point for co-regulator proteins and often contains residues that, when phosphorylated, affect the activity of the receptor. Following the AF-1 domain is the DBD that consists of two zinc-finger motifs, one of which fits the major groove of DNA and the other that is positioned above and perpendicular to it (Fig. 4). The variability in this domain translates into variability in the DNA-binding sites of NRs. Two receptors (DAX-1 and SHP-1) lack this DBD and function mainly as dominant negative repressors of other NRs.
The DBD is connected to the LBD by a flexible hinge region, which in some cases may contain a nuclear localization signal, or participate in co-regulator interactions. If viewed in three dimensions, the LBD of NRs is a three layered α-helical sandwich structure that has a ligand-binding cavity at the bottom of the domain (Folkertsma et al., 2005). For some NRs, for which no ligand has been characterized, amino acid side chains in fact fill this cavity. On the other extreme, other NRs are able to fit a variety of ligands of different size, volume and shape. Their cavity size can reach over 1000 Å³ in size, PPARs even have a cavity of 1400 Å³ in size. The LBD should also not be considered to be static as in a few highlighted cases, the LBD has been shown to fit ligands by undergoing structural changes that accommodate the new ligand (evaluated for VDR ligands in Molnár et al., 2006).

The last and most C-terminal α-helix of NRs is also known as the activation function 2 (AF-2), or helix 12 (H12) and its positioning has a crucial role in the transition between inactive and active state of the receptor. A negatively charged glutamate residue located on H12 needs to be fixed at a defined distance from a positively charged residue in H3. This fits the CoA molecule peptide (description of CoAs follows below) in size and charge and thus enables a tight intermolecular interaction to occur between the receptor in its active state and the CoA molecule (Li et al., 2003b). This important helix is highlighted in the crystal structure of the PPAR-RXR heterodimer LBDs shown in Fig. 5.
Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes

Structurally, the VDR differs from most other NRs by having a truncated N-terminus that lacks the AF-1 domain. Compared to the PPARs, VDR also has a smaller ligand-binding cavity (697 Å³), but binds its natural ligand with much higher affinity ($K_d = 1-2 \times 10^{-10}$ M). This classifies the VDR amongst the endocrine receptors, such as progesterone receptor (PR). However, the other classifications, shown in Fig. 3, based on sequence similarities, or expression profiles, classifies the receptor together with receptors, such as constitutive androstane receptor (CAR) and pregnane X receptor (PXR), classically known for their function in xenobiotic metabolism. VDR, CAR and PXR also have overlap in DNA-binding site recognition. To date, six crystal structures of VDR exist in the PDB that differ in terms of bound ligand (reference codes 1db1, 1ie8, 1ie9, 1s0z, 1s19 and 1txi). The VDR structure has the general NR topology and consists of 13 α-helices and a three-stranded β-sheet, its closest structural homolog is that of RARγ (Rochel et al. 2000; Renaud et al., 1995). The crucial H12 is in the active conformation in these ligand-bound holo-VDR structures and contacts the ligand as well as additional stabilizing residues (T415, L417, V418, L419, F422, D232, V234, I238, Q239, A267, I268, H397, Y401 and S235).

Two structurally different protein isoforms exist for PPARγ, $\gamma_1$ and $\gamma_2$, that differ in the length of the N-terminal AF-1 domain (Mukherjee et al., 1997). In total, six alternatively spliced transcript isoforms have been isolated in humans (Chen et al., 2006). It appears that the regulation of the
Review of Literature

isoforms differs. However, there is insufficient knowledge about their functional differences. PPARγ2 is predominantly expressed in adipose tissue and was activated in a similar way as PPARγ1. Both PPARα and PPARγ AF-1 domains can be phosphorylated (by phosphorylation pathways, such as the MAPK signaling cascade). Interestingly this can lead to opposite effects: activity increases have been reported for both receptors with several signaling cascades, whereas the phosphorylation of serine 112 in PPARγ was shown to decrease ligand-binding (Juge-Aubry et al., 1999; Shao et al., 1998). The latter effect suggests that intra-molecular communication exists between the different domains, an observation that will later be further evaluated in connection with how ligand-binding changes receptor functionality.

A much larger set of PPAR crystal structures is available for examination from the PDB. This set contains two unliganded structures (2gwx of PPARβ/δ and 1prg of PPARγ). Ligand-bound receptors together with a CoA-derived peptide include one PPARα structure (1k7l) and six PPARγ structures (2prg, 1wm0, 1k74, 1fm6, 1fm9 and 1RDT). One structure is crystallized together with an antagonist and a CoR-derived peptide, the PPARα structure 1kkq. The structure 1PRG was one of the first structures to be published: similar to VDR, 13 α-helices are visible; the β-strand, however, is small and four-stranded. The top part of the LBD has a rigid structure that enables the existence of a large ligand-binding pocket (rosiglitazone occupies approximately 40% of the cavity). A direct contact between the ligand and H12 is also observed (Nolte et al., 1998).

2.4.2 Ligands

The interest in the therapeutic use of 1α,25-dihydroxyvitamin D3 (1α,25(OH)2D3), the natural ligand for the VDR, has motivated the development of several analogs that maintain the beneficial roles of the natural ligand, 1α,25(OH)2D3, in controlling inflammation or cell proliferation, but lack the calcemic side effects (hypercalcemia and hypercalciuria) (Bikle, 1992; Binderup et al., 1991; van den Bemd et al., 2000). 1α,25(OH)2D3 analogs in clinical use are reviewed by Brown and colleagues (Brown, 2001). Structurally 1α,25(OH)2D3 analogs resemble the natural ligand (Fig. 6).

PPARs are activated by ligand concentrations in the micromolar range (Forman et al., 1996; Xu et al., 1999). These ligands differ significantly in their structure. To illustrate this, some of the different ligands for PPARs are shown in Fig. 6. Their potential for therapeutic use in metabolic disease is reviewed in (Berger et al., 2005). The possibility that different ligand structures may induce different CoA association profiles with more cell type-specific responses, is actively explored (a similar trend was seen in the development of 1α,25(OH)2D3 analogs). PPARα binds a diverse set of ligands. Among the most potent endogenous ligands are arachidonic acid derivatives, such as leukotriene B4, or prostaglandins (PGs) D1 and D2, whereas the pharmacological activation of the receptor goes via fibrates, such as fenofibrate or bezafibrate (Devchand et al., 1996; Forman et al., 1997). However, the endogenous activation of the receptor may mainly occur with high levels of free fatty acids that occur under fasting conditions (Kersten et al., 1999). The best-known PPARγ ligands, such as rosiglitazone, belong to the family of TZDs. Natural activation has been demonstrated by various fatty acids, arachidonic acid derivatives and their metabolites, but the inductions are weak. There is some
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes

preference for polyunsaturated fatty acids (Huang et al., 1999; Nagy et al., 1998). Initially, PPARβ/δ was less intensively studied. Systematic screens have identified saturated and unsaturated fats as ligands, but only after the development of the synthetic ligand GW501516 and observations of its positive effect on cholesterol transport in rhesus monkeys more attention has turned to this subtype (Oliver et al., 2001). Recently, the use of PPAR ligands has expanded from treatment of metabolic diseases and is being evaluated for the treatment of inflammation, hyperproliferation and cancer (Friedmann et al., 2005; Panigrahy et al., 2005).

Fig. 6. VDR and PPAR ligands. The natural ligand for VDR, 1α,25(OH)2D3, has a typical cholesterol backbone. A selection of VDR analogues with specific modifications to this structure are shown on the right (Posner, 2002). The larger ligand-binding pocket of PPARs is able to bind a diversity of ligands. The structures of the synthetic PPAR ligands GW7647 (PPARα), rosiglitazone (PPARγ) and GW501516 (PPARβ/δ) shown reflects this structural diversity. From natural PPAR activators, palmitic acid, leukotriene B4 and eicosapentanoic acid are shown.
2.4.3 NR dimerization interfaces and DNA-binding

The DBDs of NRs contain zinc atoms classifying them into the zinc co-ordinating group of DNA-binding proteins. The NR DNA-binding sites, also known as REs, are composed of hexameric binding motifs, for which an ideal (or consensus) sequence can be described. In detail, structural analysis shows that one α-helix located at the C-terminal of the first zinc finger, known as the recognition helix or P box, contacts the hexameric core binding motif in the major groove of DNA (Fig. 4) (Luisi et al., 1991). Predominantly within the group of orphan receptors, there are NRs that bind as monomers with specific flanking nucleotide preferences to RGKTCAs-type sequences (R = A or G, K = G or T). Some orphans and in addition steroid receptors, such as androgen receptor (AR), recognize two hexameric motifs as homodimers. In this form, steroid receptors recognize an inverted repeat (IR) orientation with 1 or 3 nucleotide spacing. Additionally, the steroid receptors form an exception to the RGKTCAs-rule, with a preference for RGAACA-like motifs. The remainder of receptors that include many of those sensing micro- and macronutrients and xenobiotics form heterodimers preferentially with RXR α, β, or γ. Most heterodimers recognize two hexameric motifs of RGKTCAs-type motifs in a direct repeat orientation, but also everted repeats (ERs) have been identified (Fig. 7). Conversely, the RXR-farnesoid X receptor (FXR) heterodimer is the only one known to recognize IRs. Each heterodimer has individual preferences to the nucleotide composition (including consensus sequence variations tolerated), spacing and orientation, however, overlaps exist (Rastinejad et al., 1995).

![Schematic diagram of different hexamer orientations of NR binding sites](image)

**Fig. 7. Schematic diagram of different hexamer orientations of NR binding sites.** Orientation and spacing are crucial determinants of NR binding site specificity. The DR motif has two hexamers formed of AGGTCA motifs in parallel orientation, whereas in ER and IR one of the hexamers has this motif in the reverse orientation.
The dimerization interfaces of homo- and heterodimers involve both the DBD and the LBD residues. The DBD interface follows the zinc-fingers in structure, whereas a hydrophobic region at the N-terminus of H10/11 dominates the LBD dimerization interface. In some cases, the dimerization with liganded RXR is sufficient to form a transcriptionally active complex. This effect, referred to as phantom ligand effect, classifies some heterodimers as permissive dimers. The effect that the binding of a substrate/ligand on one domain has on the properties of other domains, is referred to as allosteric regulation. Shulman and co-workers reported a statistical method to discover the energetically coupled residues within NR protein structures (Shulman et al., 2004). Interestingly, they were able to show critical residues for this allosteric network that regulate the specificity of ligand regulation. Metabolic versus endocrine ligands for VDR were shown to be differentially dependent on specific amino acid residues. In addition, their study offered a structural explanation for the phantom ligand effect of permissive heterodimers. The effect of dimerization with liganded RXRs was shown to mediate intermolecular changes via key residues that permit an activated state of the partner receptor without the presence of its ligand.

PPARs belong to the group of NRs that regulate transcription as heterodimers with RXRs (via DR1-type REs). The receptors bind the REs with PPARs occupying the 5'-hexamer, which deviates from the binding mode of most other NR-RXR heterodimers. The dimerization is also atypical concerning symmetry: the PPAR-RXR dimer is asymmetrical and residues from H10 and some from H7 and H9 contribute to the dimerization interface. In addition to DR1-type REs, binding of PPAR-RXR heterodimers to IR3- and DR2-type REs has been reported (Hsu et al., 1998; Keller et al., 1995). VDR, like the PPARs, forms heterodimers with RXRs and the dimer binds a dual hexameric RGKTCA motif (also variation to this consensus motif are recognized). As with most RXR heterodimers, the RXR molecule occupies the 5'-hexameric site. Compared to the PPARs, VDR is more flexible concerning spacing and orientation of the two hexameric sites. Classical 1α,25(OH)2D3 REs (VDREs) are DR3-type REs, however, the VDR-RXR heterodimer can also bind DR4-type REs and a number of ERs. Based on in vitro studies, ER6- to ER11-type REs can function as VDREs, however, so far only natural ER6-, ER8- and ER9-type elements have been described (Schräder et al., 1995, Tavera-Mendoza et al., 2006; Thompson et al., 2002). VDR is also able to form homodimers capable of binding to DR6-type REs (Polly et al., 1996). It is rather unclear, however, to what extent these homodimers contribute to gene transactivation since the preferred dimerization partners for VDR, the RXRs, are ubiquitously expressed.

2.5 Transactivation mediated by NRs

2.5.1 Direct protein-protein interactions

Conceptually, there are two types of co-regulator proteins that come in direct contact with the NRs, CoAs and CoRs. CoAs, such as the p160 family members SRC-1, TIF-2 or RAC3, bind the receptor in its active state, when the position of H12 and the two charged residues allow it. CoAs in turn bind
and recruit histone acetyl transferases (HATs) that relax the chromatin state locally by adding acetyl groups to the tails of histone molecules. The interaction motif of many CoAs consists of a conserved LXXLL motif (L = leucine and X = any amino acid). The leucines fit the hydrophobic pocket between the charge clamp residues, further stabilizing the interaction provided by flanking charged residues (Heery et al., 1997). This kind of interaction mode has been described also for the PPARγ CoA 1α (PGC-1α) CoA, though it may also utilize a slightly different interaction surface (Wu et al., 2003). The CREB-binding protein (CBP) and E1A-binding protein (p300) are highly related proteins that also possess the capability to interact with activated NRs (they also interact with a variety of TFs) and furthermore they may also interact with the p160 proteins. p300/CBP contains a domain with HAT-activity on its own and additionally can recruit other HATs, such as p300/CBP-associated factor (Demarest et al., 2002). Currently, proteins that possess other functionalities besides histone acetylation are considered as NR CoAs. A large number of approximately 200 CoA proteins have been described, with roles in methylation, transcriptional elongation and ubiquitination (Lonard and O'Malley, 2006). The concept of multi-subunit CoA complexes has emerged, where the collection of functions that carry out the modifications required for gene transactivation, are mediated by different proteins in the complex.

Induction of transcription requires the formation of a preinitiation complex (PIC), which includes the TFIIA to TFIIF complexes and RNA polymerase II itself (Berk, 1999). The complex initiates from the binding of the TATA-box-binding protein, followed by the TFIIA proteins A and B and the polymerase. A large multi-subunit mediator complex bridges the distance between the TF and the PIC and stimulates the phosphorylation of the largest subunit of the polymerase, initiating elongation (Woychik and Hampsey, 2002). TRAP220, a member of the DRIP/TRAP family is found as a component of the mediator protein complex and can bind NRs in the active state, utilizing the same binding mode as the p160 proteins. The initiation of this protein interaction, leads to transcriptional activity on a gene’s promoter.

CoRs bind inactive NR by utilizing, in part, the same binding interface as CoAs thus making their binding mutually exclusive. When bound to NRs, they recruit enzymes with histone deacetylase activity (HDACs) that reverse the covalent modifications caused by the HATs and thus reverse the effects of the latter enzymes. In effect, DNA and histones become more attracted. The best-characterized CoRs are NCoR1 and SMRT. These CoRs have a similar but longer receptor interaction domain than CoAs with the sequence LXXI/H/IXXXI/L (I = isoleucine, H = histidine and X = any residue). As with CoAs, an expansion in the set of known CoRs has occurred. These include the proteins small ubiquitous nuclear CoR, SHARP (interacts with SMRT), nuclear receptor interacting protein 1, Alien and scaffold attachment factor B1 that do not share sequence homology to the NCoRs (Debril et al., 2005; Dressel et al., 1999; Lee et al., 1998; Shi et al., 2001; Zamir et al., 1997).

In summary, these interactions enable NRs to come into contact with chromatin state modifiers via sequential recruitment of various protein complexes and to bridge to the basal transcriptional machinery via the mediator complex.
2.5.2 Regulation of transcription rates

The final outcome, in terms of a target gene’s mRNA transcription rate, can occur in both directions. So far, gene activation is better understood than gene repression, even though both have been observed to occur with approximately equal numbers for different NR ligands. The activation of gene transcription for ligand-activated receptors is described as a two-step process, where the switch from chromatin loosening HAT recruitment is thought to precede the recruitment of the mediator complex that bridges the NR to the RNA polymerase II at the TSS. To understand the sequence of events taking place on a target gene promoter, the well-known VDR target gene cytochrome P450 (CYP) 24 was analyzed by Väisänen and colleagues using chromatin immunoprecipitation (ChIP) assays covering a region of 10 kB upstream from the TSS (Väisänen et al., 2004). On this promoter a clear change from inactive chromatin to active chromatin (manifested by acetylated histone H4 appearance) was observed upon VDR recruitment to the promoter and at later times points mediator proteins and phosphorylated RNA polymerase II (p-PolII) were detected. However, at actively transcribed gene promoters the histone modification step may be not necessary due to a pre-existing permissive chromatin status indicated by high histone acetylation level (Sinkkonen et al., 2005).

For repression no general model exists, but instead different models have been described on a single gene level. In the case of the gene apolipoprotein (APO) C3, a replacement model was proposed, where the NR hepatocyte nuclear factor α (HNF-4α), participating in the maintenance of the basal transcriptional level of the gene, is substituted for the ligand-bound PPARs. This leads to a decrease in the transcription level due to differences in the level of transactivation mediated by these NRs. Overlapping binding sites and subsequent competition may also lead to repression, as was observed for the rat bone sialoprotein gene promoter, where the TATA box-binding protein has an overlapping binding site with the VDR (Kim et al., 1996). When occupied by VDR, gene transcription is prevented. A similar overlapping binding site is found in the rat osteocalcin gene, where an AP-1 binding site co-locates with a VDRE (Demay et al., 1992). Gene repression involving AP-1 binding sites was recently described as an essential element for late down-regulation effect mediated by the estrogen receptor α (ERα) (Carroll et al., 2006).

2.5.3 Dynamics of gene regulation by NRs

Métivier and co-workers have addressed the dynamics of NR-mediated transactivation so far in highest detail in their study (Métivier et al., 2003). This study focused on ERα recruitment to the pS2 gene promoter, where an elegant time series ChIP analysis was performed to evaluate the sequential recruitment of the different protein complexes. The results revealed a steady cyclical pattern of ERα recruitment to the promoter that was subsequently followed by proteasome-mediated receptor degradation. An initial cycle of ERα recruitment, which resulted in no mRNA production, involved the recruitment of nucleosome organization modeling SWI/SNF complex. This was followed by further histone modifications achieved via the recruitment of p160 or p300/CBP during each following cycle. The TRAP220-mediator complex followed the acetylation state modifiers and a subsequent phosphorylation of RNA polymerase II was observed. At the end of each cycle, ERα was...
directed to proteosomal degradation via AAA ATPase proteins independent of 20S or API31, a
subunit of the 20S proteasome complex. This clearance was followed by histone deacetylation prior
to the initiation of following cycles.

2.6 The role of PPARs and VDR on a genome wide scale

A global analysis of the anatomical and circadian NR expression patterns was recently carried out in
adult male mouse tissues derived from two inbred strains (Bookout et al., 2006). In this study, shared
NR gene expression patterns were hypothesized to reflect related function and common regulation.
This analysis classified the NR family in two large subgroups: NRs that function in 1) reproduction
and the central nervous system and 2) nutrient metabolism and immunity (Fig. 3). In some ways these
results reflected other classification patterns. In this study, VDR clustered together with PXR, CAR, its
two closest evolutionary relatives, and FXR, into the xenobiotic and bile acid metabolism cluster.
The highest expression of VDR is seen in tissues that play a role in calcium homeostasis, small
intestine, kidneys and bone, fitting with its classical role in bone formation, manifested in the
deficiency disease rickets. The dominancy of this feature in its function was perhaps surprising, as the
role of the receptor has been characterized in various other processes. However, supporting a more
ubiquitous role, lower expression levels were detected in nearly all mouse tissues explored.

Gene expression of the three PPAR subtype genes were also reported in various tissues: PPARβ/δ
can be found nearly in every tissue tested, whereas PPARα was expressed at highest levels in liver,
heart, kidney, muscle, intestine and brown fat, all characterized by a high level of fatty acid
catabolism. PPARγ was found in highest levels in adipocytes, in accordance with its important role in
adipocyte differentiation. Furthermore, PPARα and β/δ were clustered together with five orphan NRs
and RXRα in a metabolism cluster, while PPARγ separated to the immunity cluster, together with the
glucocorticoid receptor (GR) and the liver X receptor (LXR) α. The physiology of each receptor is
further explored below in connection to the physiological processes, in which their role has been
established. The physiological role of TFs is manifested through their target genes and accordingly,
examples of known target genes are mentioned for those processes characterized in this detail.
Additionally, some physiological effects that have been characterized from in vivo models (typically
mouse models) are also discussed.

2.6.1 Characterized physiology and known target genes of the VDR

Calcium homeostasis

In vertebrates, maintenance of bone integrity is dependent on the regulated absorption and
distribution of bone forming minerals. In small intestine and kidney, VDR target genes include those
involved in the absorption of calcium. This metal ion is one of the major components of bone
(together with phosphate) and requires facilitated absorption from the intestine. The target genes in
these tissues facilitate intracellular transport, for example, calbindin D-9K and calbindin D-28K and
transfer across membranes, such as the calcium pump PMCA (Bouillon et al., 2003). In bones, two
types of cells participate in bone remodeling: VDR expressing osteoblasts that form bone and bone-resorbing osteoclasts, which do not contain VDR. Among the genes regulated in osteoblasts are well-known VDR target genes, *osteocalcin*, associated with the mineralized matrix of bone and *osteopontin*. While the latter has an established role in bone where it binds tightly hydroxyapatite and enables osteoclast adhesion (Reinholt et al., 1990), it has also been shown to be a constitutive component of elastic fibers in skin and aorta (Baccarani-Conti et al., 1995) and bind to CD44 of T-lymphocytes modulating immune reactions (Weber et al., 1996). Role of VDR in immune reactions is further discussed later. Additionally, effects on osteoclasts are mediated indirectly via the regulation of secreted proteins in osteoblasts, such as the *osteoclast differentiation factor* that activates osteoclasts.

**Cellular proliferation and differentiation**

The first evidence of beneficial effects of VDR in hyper-proliferative disease came in the early 1980s, as it was discovered that nanomolar doses of 1α,25(OH)₂D₃ arrested the proliferation of murine myeloid leukemia cells and induced their differentiation into macrophages (Abe et al., 1981). Since then the role of VDR and the potential use of VDR analogs has been studied in various cancer models and in the modulation of cellular differentiation. Amongst other techniques, the following microarray studies have explored expression changes in response to 1α,25(OH)₂D₃ (Moll et al., 2006; Palmer et al., 2003; Swami et al., 2003; Wang et al., 2005).

The genes responsible for the regulation of cell cycle progression are obvious candidates to mediate the anti-proliferative effects of this hormone in cells. The cell cycle is tightly controlled and special checkpoints for the progression to the next step exist to prevent premature or undesired progression in cell division. On the molecular level, the different checkpoints translate into the sequential activation of cyclin-dependent kinases (CDKs). The critical cyclin-kinase complex for the initiation of cell division is the CDK4/6 complex that binds cyclin D. The activation of the kinase complex leads to the phosphorylation and inactivation of the retinoblastoma protein (Rb). This in turn releases the TF E2F from the embrace of its inhibitor, Rb and initiates the transcription of genes needed for the progression into cell division (M phase). For the inhibition of the activity of the CDK4/6 complex, there exists the specific INK4 family of CDK-inhibitors. This family comprises of the proteins p16INK4A, p15INK4B, p18INK4C and p19INK4D. In addition, members of the CIP/KIP family of CDKIs are able to inhibit all CDK complexes. This family comprises the proteins p21CIP1/WAF1, p27KIP1 and p57KIP2. Both p21CIP1/WAF1 and p27KIP1 gene expression levels have been reported to increase by 1α,25(OH)₂D₃ treatment (Liu et al., 1996; Wang et al., 1996). The former study also reported a VDRE within the human *p21CIP1/WAF1* promoter. More recently this phenomenon was studied more intensely with ChIP promoter screening for VDR binding sites. Together, these studies demonstrate that the up-regulation of the *p21CIP1/WAF1* gene is directly mediated by the VDR and involves several binding sites and synergy with p53, another TF that is involved in cell cycle regulation (Saramäki et al., 2006). In addition, the induction of *p21CIP1/WAF1* is also reported to occur via the IGFBP-3 gene, another direct target of VDR (Peng et al., 2004). This suggests both primary and secondary roles for 1α,25(OH)₂D₃ and its receptor, VDR, in the regulation of this gene. In addition, the expression levels
of several other CDKIs, cyclins and CDKs, are regulated by VDR (Jensen et al., 2001). Moreover, VDR has been shown to directly regulate a gene that encodes for a protein that is involved in cellular DNA repair mechanisms, GADD45 (Jiang et al., 2003). This protects from UVB-induced DNA damage that is critical, especially for skin that is the most exposed tissue to this type of electromagnetic radiation (De Haes et al., 2005). Skin is also able to synthesize 1α,25(OH)2D3, when exposed to UV-light, and is therefore our body’s main way of producing this hormone.

In addition to cell cycle checkpoint genes, VDR has a role in the regulation of proto-oncogenes and tumor suppressor genes. For example, the oncogenic TF c-fos gene transcription (Mathiasen et al., 1993) has been reported to be rapidly up-regulated in several cell lines and among tumor suppressor proteins, the regulation of E-cadherin has been shown to be transcriptional regulated by 1α,25(OH)2D3 (Pålmer et al., 2001). Further, the modulation of growth factor signaling, such as that of the tumor growth factor (TGF) β and insulin-like growth factor (IGF)-I signaling, plays a role in the anti-proliferative/pro-differentiation effects of the VDR. For instance, a VDRE has been identified from the TGFβ2 gene promoter (Wu et al., 1999) and regulation of TGFβ receptor-1 has been observed with several microarrays (Pålmer et al., 2003). Several members of the gene family that encode the IGFBP proteins, including the IGFBP-3 gene, whose promoter contains a VDRE (Drivdahl et al., 1995). These proteins can bind and sequester the circulating IGFs making its growth-promoting stimulus inaccessible to cells (Baxter, 1994). On the other hand, membrane bound IGFBPs have a lowered affinity for IGFs and function to present the growth factor to its receptor (Kelley et al., 1996). Such an effect was observed for IGFBP-5 in bone fractures, where it functioned to locally increase the concentration of IGF-1 promoting bone reformation (Mohan et al., 1995).

**Immune reactions**

Several cells of the immune system, such as monocytes, antigen presenting macrophages, dendritic cells and activated T-lymphocytes, express the VDR. An interaction between VDR and AP-1 signaling was suggested as one underlying mechanism for the immunosuppressive action of VDR that leads to suppression of inflammatory cytokine expression (Alroy et al., 1995). VDR was shown to block an interaction of the TFs Fos, Jun and NF-AT (Towers et al., 1999). Due to their immunosuppressive ability, VDR ligands have been found beneficial in several autoimmune diseases, such as multiple sclerosis where an autoimmune reaction is triggered against self-antigens in the brain, rheumatoid arthritis where inflammation is triggered in the joints, in inflammatory bowel disease where the gastrointestinal tract is affected, and in type I diabetes where the pancreatic insulin producing cells, β-cells, are affected (the immunological functions of the 1α,25(OH)2D3 endocrine systems are reviewed in (Hayes et al., 2003)). In all these diseases evidence for the beneficial effects of 1α,25(OH)2D3 has accumulated.

Multiple sclerosis develops when an immune response is triggered by the inappropriate recognition of self-epitopes on myelinated nerve fibers. This leads to the recruitment of T cells and macrophages into the central nervous system resulting finally in demyelination of nerves that leads to the...
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing *in silico* methods into the analysis of nuclear receptor regulomes

Symptoms of the disease. The differentiation of T helper 1 (Th1)-type cells is controlled by antigen stimulation and cytokines, particularly interleukins (IL) -12 and -23. This cytokine signaling translates to Th1-specific gene expression profile via the TF T-bet (Szabo et al., 2000). VDR has been shown to inhibit IL-12 production directly leading to the suppression of Th1-response (D’Ambrosio et al., 1998). The Th2 population, functioning in defense against parasites and hypersensitivity reactions, is not affected. Mattner and colleagues addressed this in a mouse model immunized with self-antigen peptide and treated with 1α,25(OH)2D3. Inflammatory infiltration and demyelination of brain and spinal cord were reduced (Mattner et al., 2000). Effects mediated via inhibition of IL-12 signaling have also been proposed to underline the beneficial effects of VDR in type 1 diabetes (Gregori et al., 2002). Among other signaling pathways, the up-regulation of TGFβ and IL-4 have been proposed to contribute to the effect (Cantorna et al., 1998). In the treatment of rheumatoid arthritis, convincing clinical evidence exists to support the beneficial, anti-inflammatory effects of 1α,25(OH)2D3 (Andjelkovic et al., 1999). This disease is one of the most common chronic inflammatory diseases and results in tissue damage of the joints involving the matrix modeling proteases. In addition, prostaglandin signaling plays a role in the immune system and inflammatory processes associated with rheumatoid arthritis. Interestingly, IL-1β-stimulated prostaglandin E2 synthesis is completely inhibited by 1α,25(OH)2D3 treatment and also metalloproteinases are regulated (Tetlow et al., 1999).

Antigen-presenting cells, in particular dendritic cells, are also affected by the immunosuppressive activity of 1α,25(OH)2D3. These cells play a central role in regulating immune response to self and foreign antigens. During the normal immune response, T cell response and specificity are conferred through the clonal restricted T cell receptor, which recognizes major histocompatibility complex class I and class II molecule complexed with peptides. However, the potency of this activation also depends on the maturation status of the dendritic cells, in particular the display of costimulatory surface proteins (Banchereau, Steinman, 1998). Important evidence of 1α,25(OH)2D3 effect on antigen presentation comes from VDR knockout mouse studies, where dendritic cells from VDR-deficient mice showed a significantly higher level of maturation markers such as class II MHC, CD40, CD80 and CD86 on cell surface. In the presence of GM-CSF and IL-4, 1α,25(OH)2D3 completely inhibited DC differentiation manifested by the low-level expression of IL-12 and maturation markers (Griffin et al., 2001). This effect has lead to the use of 1α,25(OH)2D3 or its analogs in graft rejection therapies. Furthermore, recent evidence also suggests that 1α,25(OH)2D3 treatment leads to the induction of tolerogenic dendritic cells, which protect the allograft rejection (Adorini et al., 2003).
2.6.2 Characterized physiology and target genes of PPARs

2.6.2.1 PPARα

**Fatty acid oxidation**

The role of PPARα is best established in fatty acid oxidation, which takes place predominately in the liver. An extensive list of direct targets, for which binding sites have been identified exists for mouse PPARα, however, not all have a characterized PPRE in the human gene (reviewed in detail in (Mandard et al., 2004)). The use of fatty acids in energy production requires the uptake of free fatty acids and fatty acids carried by lipoproteins from the circulation into liver cells. The *fatty-acid transport protein-1* gene is a known PPARα target in mouse and the *CD36* gene (APO receptor) has been described as a target in both species (Frohnert et al., 1999; Sato et al., 2002). The *hepatic fatty acid binding protein*, responsible for the intracellular trafficking of fatty acids in liver cells has a characterized binding site in the mouse gene (Poirier et al., 2001). Its role may be in enhancing the signaling role of fatty acids, by mediating their traffic into the nucleus, where they may activate PPARs in a feed-forward fashion to increase the transcriptional levels of genes that participate in the subsequent steps of fat metabolism.

Prior to the translocation of fatty acids to the mitochondria, they need to be activated by the enzyme encoded by the gene *long-chain fatty acyl-CoA synthetase*, for which a PPRE was described in the rat promoter (Schoonjans et al., 1995). The cascade of human PPAR targets continues with the mitochondrial translocase genes *carnitine palmitoyl acetyltransferase (CPT) 1α, CPT1β* in the outer membrane and *CPT2* in the inner membrane (Barrero et al., 2003; Mascaro et al., 1998; Napal et al., 2005). The enzymes functioning inside the mitochondria in the first steps of β-oxidation, such as the respective *acyl-CoA-dehydrogenase* genes for the different length of fatty acids, were described in mice as direct PPAR target genes (Gullick et al., 1994) and have also been shown regulated in humans by microarrays (Vanden Heuvel et al., 2003). The first identified direct PPARα target gene, *acyl-CoA oxidase (ACOX1)*, encodes the enzyme catalyzing the first step of peroxisomal β-oxidation (Tugwood et al., 1992). The orthologous human gene was also later characterized as being a PPAR target (Varanasi et al., 1996). Besides mitochondria and peroxisomes, the smooth endoplasmic reticulum also participates in fatty acid oxidation, but at a minor level. This is the site of ω-oxidation and the enzymes catalyzing these reactions, belonging to the 4A/F group of CYP enzyme, were characterized as direct target genes PPARs in rodents (Johnson et al., 1996).

**Other metabolic pathways regulated in the liver**

When fatty acids serve as the only energy source for a prolonged time, the liver starts gluconeogenesis in an attempt to maintain blood glucose levels and later, to synthesize ketone bodies, that go on to serve the energy needs of the brain. PPARs also participate in the regulation of genes in the gluconeogenesis and ketone synthesis pathways. If these conditions persist, amino acids become the main energy source of liver cells and this shift is also regulated. Liver is also the main site of biotransformation reactions that inactivate xenobiotics and some endogenous compounds. The role of
PPARα in the regulation of genes participating in these reactions has been well reviewed in the literature and will not be discussed here (Mandard et al., 2004).

**Regulation of lipoprotein levels**

The clinically most interesting aspect of gene regulation by PPARα is the regulation of lipoprotein metabolism. In this pathway, most of the genes are also known to be direct targets of PPARs in humans. The fibrate drugs that activate PPARα have been shown to increase high density lipoproteins (HDL) level in plasma and decrease the triglyceride levels. Additionally, in physiological fasting conditions, the transport of fatty acids to the liver by the carrier proteins is enhanced. The genes APOA1 and APOA2, the two major components of HDLs, are direct targets in humans (Vu-Dac et al., 1994; Vu-Dac et al., 1995). The HDL surface proteins, encoded by the genes, phospholipid transfer protein (PLTP) and the gene encoding the receptor for HDL in liver tissue, the scavenger receptor-class B type 1 (SR-BI), were found to be direct targets of PPARs (Malerød et al., 2003; Tu and Albers, 2001). In connection with triglyceride levels, the down regulation of the APOC3 has been described to be PPAR-dependent (Hertz et al., 1995). APOC3 functions to suppress the lipoprotein lipase (LPL) gene. This suppression ultimately leads to elevated triglyceride levels. Furthermore, levels of LPL protein have been shown to be up-regulated by PPARs, further substantiating the effect (Li et al., 2002). In addition, the human APOA5 gene, which is involved in the maintenance of normal triglyceride levels, is a direct PPAR target (Prieur et al., 2003).

**Other pathways**

In other tissues, a role for PPARα has been described in the regulation of uncoupling proteins (UCPs). The thermoregulatory UCP1 gene, in brown adipose tissue (in mice, adult humans lack brown adipose tissue), UCP2 that participates in insulin production in pancreatic β-cells and UCP3, expressed mainly in skeletal muscles (Armstrong and Towle, 2001; Barberá et al., 2001; Brun et al., 1999) have all been described as PPAR target genes. Additionally, a role in bile acid absorption for PPARα was recently described by regulating the ileal fatty acid binding protein (FABP6) and the transporter solute carrier family (SLC) 10, member A2 (Landrier et al., 2005).

**2.6.2.2 PPARγ**

To appreciate the role of PPARγ, it is important to understand the metabolic crosstalk between the fat storing (adipose) and the main fat burning (liver and muscle) tissues, because directly linked to this tissue network is also the control of blood glucose levels (Evans et al., 2004). Insulin is best known for its role in the regulation of blood glucose levels, however it also regulates triglyceride catabolism via the inhibition of the hormone sensitive lipase (LIPE) gene, a known PPAR target, already suggesting a link between the regulation of the two energy sources (Holm, 2003). The key observation that links the regulation of lipid levels (mediated by PPARs) and insulin sensitivity is that lipid levels in tissues in fact have an effect on glucose homeostasis. The effect has given rise to the lipotoxicity hypothesis, which states that abnormal accumulation of triglycerides and fatty acyl-CoA in both muscle and liver cells can result in insulin resistance (Shulman, 2000). Thus the storage of
fatty acids in fat tissue, promoted by PPARγ, prevents lipotoxicity from developing. This effect, together with regulation of signal molecules secreted from the fat tissue, such as adiponectin and leptin (both having wide spread effects on whole body energy metabolism), can underline the positive effect of PPARγ ligands in the use of the treatment of diabetes (Kallen and Lazar, 1996; Maeda et al., 2001). In addition, PPARγ has been reported to play a role in atherosclerosis, where likewise the abnormal deposition of fat, in this case as cholesterol, into blood vessels causes the disease. PPARγ has been characterized to mediate the processing of cholesterol back into transport on HDL particles, explored more in detail below.

**Differentiation of adipocytes and fat storage**

The longer isoform of PPARγ, PPARγ2 (28 additional amino acid at the N-terminus), has been characterized as the master regulator of the formation of fat cells and their normal function in the adult. Much of the discussion related to the role of fat tissue in glucose homeostasis deals with this isoform. Evidence for its pivotal role in the process comes from various observations: 1) the expression levels of the PPARγ2 are increased during differentiation of fat cells, 2) non-adipogenic cells can be differentiated into fat cells by PPARγ2 overexpression and 3) PPARγ2 knockout mice (that also lack the PPARγ1 isoform of the protein as well) fail to develop fat tissue (Barak et al., 1999; Rosen et al., 1999; Tontonoz et al., 1994). The differentiation process was studied in a microarray experiment performed on human adipocytes, where PPARγ was knocked-down using an antisense oligo knock-down system (Perera et al., 2006). A total of 278 genes were found to be dependent on PPARγ for the maintenance of their normal expression (regulation absent in knock-down), out of which 77% were genes for which the expression levels differed between adipocytes and pre-adipocytes. The biological processes characteristic of fat cells appear to be enriched on the list of PPAR targets derived from this study. These include the process of lipid storage into fat from circulation mediated by PPAR target genes, such as LIPE, perilipin (Holm, 2003) and CD36. Activation of these metabolic pathways shifts the triglyceride excess from the muscle, the liver and the circulation, into adipose tissue. This results in improvements to insulin sensitivity. In addition, other signaling pathways via tumor necrosis factor α (TNFα) and resistin that promote insulin resistance were inhibited by PPARγ ligands (Hallakou et al., 1998).

**Lipid processing and signaling by macrophages**

The PPARγ1 isoform has a wider tissue distribution. Its role has been recently demonstrated in steps leading to atherosclerosis, specifically in its contribution to macrophage transition into foam cells. The immune system and inflammation reactions play a critical role in the development of this disease and PPARγ agonists have been found to regulate their effect in a way, which improves the symptoms of the disease. The development of atherosclerosis starts with the infiltration of monocytes through endothelium into the intima layer of blood vessels where they engulf oxidized low-density lipoprotein (LDL) particles and transform into foam cells. These cells secrete cytokines that promote inflammation and smooth muscle cell proliferation. The uptake of ox-LDL is mediated by the PPAR target gene CD36. In addition, PPARγ increases the expression of the transporter gene ABCA1.
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes

(although this effect has been reported to be indirect mediated by the induction of the direct target \(LXR\alpha\), which mediates cholesterol efflux from cells (Chawla et al., 2001a). To study the genes involved Hodgkinson and co-workers performed a microarray study on human macrophages that had been treated with both synthetic ligands (ciglitazone and rosiglitazone) and endogenous PPAR\(\gamma\) (15\(\delta\)-PGJ2) ligands, with a 24 h exposure time (the panel of ligands also included the PPAR\(\alpha\) ligand WY14643 and a combination of rosiglitazone with the antagonist BADGE). From this, genes involved in lipid processing were reported (Hodgkinson and Ye, 2003). Therefore, the authors proposed a model, where the sequence of events stimulated by PPAR\(\gamma\) involves the uptake of pro-atherogenic oxLDL, which gets processed and exported to HDL. This leads to an anti-atherogenic effect overall, as lipid trafficking via HDL alleviates their accumulation in blood vessels. Furthermore, the down-regulation of \(TNF\alpha\) and other cytokine genes, the products of which promote inflammation contributes to the beneficial effects observed. Thus, similar to processes regulated in fat tissue, processing and signaling events in adipocyte tissue are under the regulation of PPAR\(\gamma\)1 in macrophages. It is worth noting that the PPAR\(\alpha\) ligand stimulated the same genes, though with slightly different inductions suggesting that the PPAR isotypes may share this role that so far is mainly characterized for PPAR\(\gamma\), though the high ligand concentration used for PPAR\(\alpha\) (100 \(\mu\)M) may have also activated the other PPAR subtypes.

2.6.2.3 PPAR\(\beta/\delta\)

Energy expenditure

The interest towards PPAR\(\beta/\delta\) has grown only recently, as its beneficial effects on lipid profiles were discovered (reviewed in (Barish et al., 2006)). A mouse model overexpressing this subtype was found to induce a muscle-type switching phenotype that increases endurance, which also gave them the nickname ‘‘marathon mice’’ (Wang et al., 2004). In adipose tissue, PPAR\(\beta/\delta\) was found to increase thermogenesis (via UCPs) and fatty acid oxidation (shared target genes with PPAR\(\alpha\)). The PPAR\(\gamma\) target genes responsible for lipogenesis and lipid storage were found to be unresponsive to PPAR\(\beta/\delta\) (Wang et al., 2003a). The lean phenotype of the mice was in accordance with this. In addition, levels of HDL and triglycerides were increased in blood circulation (similar to the effect observed with PPAR\(\alpha\) ligands) and glucose production in liver was decreased due to a shift to the pentose phosphate pathway. Recently, beneficial effects were reported in human subjects (Sprecher et al., 2006).

Differentiation and wound healing in skin

Additionally, the role of PPAR\(\beta/\delta\) in skin and wound healing has been explored (Wahl, 2002). Schmuth and colleagues showed in a microarray study that PPAR\(\beta/\delta\) stimulates differentiation via the induction of keratinocyte differentiation related genes, such as involucrin, desmoplakin and cystatin A (Schmuth et al., 2004). Simultaneously, genes involved in lipid accumulation, such as adipocyte differentiation related protein (ADRP) (intronic PPRE reported in (Kersten et al., 2000)) were upregulated (these genes are shared targets...
with PPARα and PPARγ). The pro-differentiation effect is also observed in the wound healing process, as characterized with knock-out mice. PPARβ/δ is up-regulated in inflammation and stimulates differentiation of the keratinocytes and protects them from apoptosis.

The discovery of the role of PPARs in humans on whole organism level is not straightforward, as is evident from the investigations of the role of the VDR as well. In fact, a lot of valuable data concerning the physiology of PPARs has been derived from rodent models/experiments, as discussed above. However, a couple of experiments to compare the role of the orthologs have suggested that not all observations in rodents translate to the human system: Vanden Heuvel and colleagues characterized the genes responsive to PPARα ligand (WY14643, 6 h exposure) from human and rat liver cell lines (HepG2 and FaO respectively) and found only a marginal 2.5% overlap in the genes regulated by this ligand (Vanden Heuvel et al., 2003). Moreover, the mouse model, in which the mouse PPARα was replaced by the human ortholog, also revealed that the human receptor, in fact, regulated different target genes when expressed in mouse than its mouse ortholog (Cheung et al., 2004). On the other hand, several publications addressing individual target genes report observing the regulation in both species, but with different fold inductions. However, the possibility that regulation may integrate multiple signals in different species and not manifest in an identical physiological effect, should be appreciated.

Furthermore, the DBDs of the three PPAR subtypes in human are highly conserved (over 80% identical residues) and their LBDs fairly similar as well (approximately 70% identity). The similarity of the PPAR subtypes would suggest that, in fact, some target genes and regulatory pathways may indeed be under the regulation of more than one subtype but have just been described more in detail for one PPAR subtype so far. This effect was observed by Tachibana and co-workers in a cellular model system (HepG2 liver cell line) using a tet-off expression system, where the regulation of the expression of each subtype could be controlled (Tachibana et al., 2005). Several over-expression conditions were tested with ligand treatments (24 h) on subtype level and this resulted in several shared target genes showing up in the microarray results. Each PPAR was found to induce fatty acid oxidation in these cells. PPARα was found most effective, fitting well its characterized role, but the results also provided evidence for widespread sharing of target genes, when expression levels are manipulated. This suggests that to some extent also overlap in target genes is to be expected, where selectivity may depend on expression levels and ligand concentrations used. However, since the physiological roles of the receptors have been shown to differ in several studies, this divergence in vivo, from a structural perspective, could be postulated to reflect their structural differences, leading to differential recruitment of co-factors, for example. Indeed, unliganded PPARβ/δ has been shown to associate with CoRs with higher affinity than the other subtypes, also when bound to DNA (Shi et al., 2002) and the effects mediated by this subtype in muscle cells have been tightly linked with the association with PGC-1α CoA after ligand-binding (Hondares et al., 2007).
2.6.3 Overlap in VDR and PPAR signaling

Although PPARs are better known for their role in lipid level regulation, they are also known to influence cell proliferation and differentiation, processes that were discussed in connection to VDR previously (as exemplified by the keratinocyte and adipocyte differentiation discussion above). Therefore, it is not surprising that shared target genes exist. For example the gene G0/G1 switch gene 2 (G0S2) that results in growth arrest is a shared target of VDR and PPARs (Pálmer et al., 2001) (Zandbergen et al., 2005). PPARs have also been connected to the regulation of the INK4 family members, p16 was reported as a direct target recently (Gizard et al., 2005). On the other hand, VDR has recently been connected to regulation of adipocyte function. Insulin-induced gene-1 (Insig-1) and its homolog the Insig-2 gene encode closely related proteins of the endoplasmic reticulum that block proteolytic activation of sterol regulatory element binding proteins, which are membrane-bound TFs that activate synthesis of cholesterol and fatty acids. These proteins are known to restrict lipogenesis in mature adipocytes and block differentiation of pre-adipocytes. Interestingly, the Insig-1 gene is regulated by PPARγ, whilst Insig-2 gene regulation by VDR was recently reported (Kast-Woelber et al., 2004; Lee et al., 2005b). Both genes are induced during adipocyte differentiation and seem to restrict the size of this cell type (Li et al., 2003a). This is of significance, because smaller adipocytes display higher insulin-sensitivity, indicating that both NRs could influence insulin-sensitivity.

There is also a connection between osteoporosis and atherosclerosis (reviewed in detail in Hamerman, 2005). Several matrix proteins are shared between bone and arterial vessel wall, including osteopontin. Activation of PPARγ is able to re-direct the differentiation of mesenchymal progenitor cells from osteoblast precursors to adipocyte precursors, which reduces bone formation. Combined with the down-regulation of the osteoprotegerin gene, that inhibits osteoclastogenesis, these effects could lead to calcification of vessel walls seen in atherosclerosis. Under current knowledge, however, this link cannot be fully established, but these important signaling pathways may converge. The role of VDR in arterial smooth muscle cells was studied recently by Wu-Wong and collaborators (Wu-Wong et al., 2006a). In a separate study, this group also performed an evaluation of 1α,25(OH)2D3 analogues in the vessel calcification process (Wu-Wong et al., 2006b). Further understanding of VDR and PPAR signaling can thus lead to improved therapies against both of these diseases, whose incidence increase with age.

2.7 Methods for the in silico prediction of NR binding sites and target genes

For in silico analysis of NR binding sites, a couple of dedicated tools exist. Nubiscan is a WM-based tool that allows users to input their own search matrix individually for both hexameric motifs (in the case of dimeric binding NRs) (Podvinec et al., 2002). In addition, ready matrices exist for most receptors. Distinct from the WM approach employed above and distinct from approaches dedicated on specific NR recognition motifs, the NHR Scan tool allows the user to search simultaneously for all kinds of orientations and spacing, with variations allowed based on a fused list of all known NR
binding sites (Sandelin and Wasserman, 2005). The idea to make an all-inclusive search derives from the similarity of the DBDs of NRs but leaves it for the users own expertise to select from the result list the ones recognized by the NR of interest. In addition to these dedicated tools, several other search tools that use TF databases, such as TRANSFAC or JASPAR (ConSite) contain WMs for the better characterized NRs (Galperin, 2007).

The WM-approach to discover target genes was used for a genome-wide search of PPREs (Lemay and Hwang, 2006). A co-localizing motif was found frequently but the authors were unable to determine what could bind it. A different approach to benefit from a collection of known binding sites of the ERα and β was used by Lin and co-workers (Lin et al., 2004). They also included the surrounding sequence of functional binding sites with the assumption that these neighboring DNA elements may contain co-localizing motifs useful for the detection of functional binding sites. Other whole genome datasets were based on a further experimental evaluation of binding preferences to consensus variants, among others, Bourdeau and collaborators focused their search on high affinity ERα/β binding sites (Bourdeau et al., 2004) and Wang and co-authors scanned the genome for single (and zero) nucleotide consensus deviants for the VDR DR3- and ER6-type REs, for which experimental binding data was first provided (Wang et al., 2005).

Each of the whole genome datasets published for NR binding sites placed special importance on one or two parameters defining functional binding sites, namely binding strength of consensus variants, conservation, or contribution of the surrounding sequence. Often, the search window was defined to narrow down the list of predicted elements to those in the vicinity of the TSS of the respective gene. Promoter analysis done in earlier times (prior to the release of the whole genome) started with cloning of promoter fragments upstream of the TSS and scanning functional regions by mutational analysis. In total, this has yielded a vast list of published proximal elements. With the sequence data at hand, though, evidence against a positional bias is increasing. Additionally, more distal upstream NR binding sites have been described and, in addition, functional sites within the transcribed region, particularly within the introns of genes have been found. The possibility of DNA loops within chromatin units serves as an explanation of how such distal sites can come into contact with the TSS. Based on a comparative genomic analysis, this is a feature that applies to NRs. A study by Xie and colleagues analyzed multiple species to discover conserved TF binding sites and, interestingly, for certain TFs a positional bias for the proximal promoter was indeed found (Xie et al., 2005). However, there were also exceptions to the rule, such as NR DNA-binding motifs that were found with equal distribution throughout the analyzed +/-5 kb around the TSS, suggesting that location is not a functional constraint for NR binding sites. Each of the variables, such as requirements for binding strength, location, conservation and surrounding sequence, places different limits on the effectiveness of NR RE identification by in silico methods, but a consensus for their most efficient application is lacking.
3. Aims of the Study

PPARs belong to the low ligand affinity NR group. Some structure-function relationships in the activation of these receptors involve major areas of uncertainty. In addition, the characterization and prediction of their target genes and those of other NRs including VDR has not been studied in sufficient detail (in terms of binding site composition) to result in reliable in silico models to characterize their regulomes, as explained in the literature review.

Therefore the specific aims of this study were:

1. To investigate the mechanism of ligand-independent CoA interaction of human PPARs.

2. To characterize direct VDR and PPAR target genes of the IGFBP gene family to investigate possible overlap in their signaling.

3. To develop reliable in silico binding site prediction methods for the identification of VDREs and PPREs based on the characterization of IGFBP genes.

4. To correlate in silico binding site composition predictions with mRNA regulation.

5. To evaluate the utility of various parameters of binding site composition for target gene prediction, including location relative to TSS, conservation between species, number of putative REs and their binding strength and test this on a genomic scale.
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes
4. Materials and methods

4.1 Materials

4.1.1 Ligands

In the following sections the roman numerals refer to the four studies forming the thesis (see page 6 for numbering). NR ligands used in the studies are shown in Table 5. Stock solution of 1α,25(OH)2D3 was in 2-propanol and further dilution were made in EtOH. Other compounds were stored and diluted in DMSO. The translation inhibitor cycloheximide was dissolved in EtOH and used in a final concentration of 10 μM.

Table 5. NR ligands used in the studies

<table>
<thead>
<tr>
<th>Compound</th>
<th>NR</th>
<th>Source</th>
<th>Used in</th>
</tr>
</thead>
<tbody>
<tr>
<td>1α,25(OH)2D3</td>
<td>VDR</td>
<td>Dr. L. Binderup, LEO Pharma, Ballerup, Denmark</td>
<td>I, II</td>
</tr>
<tr>
<td>CITCO</td>
<td>CAR</td>
<td>Biomol, Copenhagen, Denmark</td>
<td>I</td>
</tr>
<tr>
<td>WY168281</td>
<td>PPARα</td>
<td>Dr. P. Honkakoski, University of Kuopio, Finland</td>
<td>I</td>
</tr>
<tr>
<td>GW7647</td>
<td>PPARα</td>
<td>Alexis Biochemicals, San Diego, CA, USA</td>
<td>III, IV</td>
</tr>
<tr>
<td>Rosiglitazone</td>
<td>PPARγ</td>
<td>Dr. M.W. Madsen, LEO Pharma, Ballerup, Denmark</td>
<td>I, III, IV</td>
</tr>
<tr>
<td>L783483</td>
<td>PPARβ/δ</td>
<td>Dr. M.W. Madsen, LEO Pharma, Ballerup, Denmark</td>
<td>I</td>
</tr>
<tr>
<td>GW501516</td>
<td>PPARβ/δ</td>
<td>Alexis Biochemicals, San Diego, CA, USA</td>
<td>III, IV</td>
</tr>
</tbody>
</table>

4.1.2 Cell lines

Cell lines used and their growth conditions are described below (Table 6). Medium was supplemented with 2 mM L-glutamine, 0.1 mg/ml streptomycin and 100 U/ml penicillin. Cells were grown in humidified 95% air / 5% CO2 incubator. When split for experiment, the cells were grown overnight in phenol-red free DMEM supplemented with 5% charcoal-stripped fetal bovine serum (FBS). FBS was stripped of lipophilic compounds by stirring it with 5% (w/v) activated charcoal (Sigma-Aldrich) for 3 h at room temperature (RT). Charcoal was then removed by centrifugation and sterile filtration.
Table 6. Growth conditions for the cell lines used in the studies

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Medium</th>
<th>FBS (%)</th>
<th>Used in</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCF-7</td>
<td>α-MEM</td>
<td>5</td>
<td>I, II</td>
</tr>
<tr>
<td>HEK293</td>
<td>DMEM</td>
<td>10</td>
<td>I, III, IV</td>
</tr>
<tr>
<td>PC-3</td>
<td>DMEM</td>
<td>10</td>
<td>II</td>
</tr>
<tr>
<td>SaOS-2</td>
<td>DMEM</td>
<td>10</td>
<td>II</td>
</tr>
<tr>
<td>HepG2</td>
<td>DMEM</td>
<td>10</td>
<td>III, IV</td>
</tr>
</tbody>
</table>

4.1.3 DNA constructs

**Protein expression constructs**

Previously sub-cloned protein expression vectors of the full-length cDNAs for human CAR (Baes et al., 1994), human VDR (Carlberg et al., 1993), human RXRα (Levin et al., 1992), human PPARα (Sher et al., 1993), human PPARγ (Schmidt et al., 1992) and human PPARβ/δ (Tontonoz et al., 1994) were used for mutagenesis, T7 RNA polymerase-driven *in vitro* transcription/translation and the overexpression of the respective proteins in mammalian cells.

The NR interaction domains of mouse SRC-1 (amino acids 597 to 791) (Onate et al., 1995), human RAC3 (amino acids 673 to 1106) (Li et al., 1997) and human TIF-2 (amino acids 646 to 926) (Voegel et al., 1996) were previously subcloned into the *BamH*I/*HindIII* cloning site of the GST fusion vector pGEX/KGK (Amersham Biosciences, Buckinghamshire, UK). The GST fusion proteins were used in supershift assays.

**Reporter gene constructs**

The core sequences of the REs or RE-containing regions are indicated in the original articles. These were fused to the *thymidine kinase* (*tk*) promoter driving the firefly luciferase gene (LUC) in a pGL3-derived Luciferase Reporter Vector (Promega, Madison, WI, USA). The *Xba*I-restriction site at nucleotide position 1742 had been removed from the original vector by mutagenesis. The individual REs were cloned as two-copy constructs of the RE into the remaining *Xba*I cloning site. For the cloning of RE-containing regions PCR primers with *Sac*I (forward primer) and *Xba*I (reverse primer) cutting sites were used in the PCR reaction with genomic DNA as a template. After restriction enzyme digestion the inserts were cloned between the *Sac*I and *Xba*I sites of the vector. All constructs were verified by sequencing.

**Site-directed mutagenesis**

Mutations produced to expression vectors are indicated in the original articles. The protein expression construct for PPARγ was used as a template for site-directed mutagenesis to result in specific amino
Materials and Methods

acids changes in the protein produced (used in I). Reporter gene constructs of RE-containing regions were used as a template for site-directed mutagenesis to result in nucleotide substitutions in the RE sequence (II, III, IV). The QuickChange™ Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA, USA) was used as described by the manufacturer together with oligonucleotides containing the mutated sequence (MWG BioTech AG, Ebersberg, Germany). Each point mutation was confirmed by sequencing.

4.2 Methods

4.2.1 In vitro methods

4.2.1.1 Transfection and luciferase reporter gene assay

Cells were seeded into six-well plates and grown overnight in cell line-specific medium (Table 5) supplemented with 5% (w/v) charcoal-treated FBS. Transfection methods used in the studies employed either N-[1-(2,3-dioleoyloxy)propyl]-N,N,N-trimethylammonium methylsulphate (DOTAP, Roth, Karlsruhe, Germany) (I, II) or polyethylenimine (PEI) (Laffitte et al.) (III, IV) to form liposomes. In the DOTAP method per each well 10 μg of DOTAP was incubated for 15 min at RT with 1 μg of the reporter plasmid and 1 μg of the pSG5-based reporter expression vector in a total volume of 100 μl. In the PEI method the incubation of DNA with the transfectant was preceded by a 15 min incubation of the DNA plasmids (1 μg each) with 50 μl of 150 mM NaCl and a simultaneous incubation of 10 μg of PEI (Sigma-Aldrich, St. Louis, MO, USA) with 50 μl of 150 mM NaCl. Subsequently, the two solutions were combined and incubated another 15 min at RT. In both methods, phenol red-free Dulbecco’s modified Eagle’s medium (DMEM) was added to a final volume of 1 ml. Medium of the cells was replaced and the transfection medium added. After 4 h ligands were added in the indicated concentrations in phenol red-free DMEM supplemented with 5% FBS.

On the next day, 16 h after ligand addition, cells were washed once with ice-cold PBS (140 mM NaCl, 2.7 mM KCl, 1.5 mM KH₂PO₄, 8.1 mM Na₂HPO₄·2H₂O) and lysed with 200 μl of reporter gene assay lysis buffer (Roche, Mannheim, Germany) per well. The cells were incubated for 15 min at RT at a rocking platform. The lysate was collected and the cell material pelleted with a 5 min centrifugation at 20,000 g at RT. To determine the total protein concentration, 10 μl of the lysate was taken and combined with 200 μl of Protein Assay buffer (Biorad, München, Germany) and absorbance measured at 595 nm. For luciferase activity measurement 150 μl of the lysate was combined with 100 μl of luciferase substrate (Luciferase Reporter Gene Assay buffer, Canberra-Packard, Groningen, Netherlands) and measured with the Highpass filter. A 96-well plate reader luminometer was used for both measurements (Anthos Labtec Instruments, Wals/Salzburg, Austria). The luciferase activities were normalized against total protein concentration.
4.2.1.2 *In vitro* transcription and translation

To produce nuclear receptor proteins for gelshift and supershift assays a coupled *in vitro* transcription/translation reaction was performed using their cDNA expression vectors described in the materials (Craig et al., 1992). Rabbit reticulocyte lysate was used as recommended by the supplier (Promega).

25 µl rabbit reticulocyte lysate, nuclease treated
1 µl amino acid mix (1 mM)
1 µl RNasin (40 U/µl, Fermentas, Vilnius, Lithuania)
10 µl rNTPs (5 mM)
3 µl MgCl₂ (50 mM)
5 µl template DNA (0.4 mg/ml)
1 µl T7 RNA polymerase (20 U/µl, Fermentas)
4 µl H₂O

The reaction was incubated for 90 min at 30ºC. Protein amounts were quantified by test translation in the presence of [S³⁵]-methionine for each protein batch produced. The concentrations of the receptor proteins were adjusted to approximately 4 ng/µl (10 ng is approximately 0.2 pmol) taking into account the number of methionine residues per protein.

4.2.1.3 Bacterial overexpression of proteins

For supershift assays GST-fusion proteins of SRC-1, RAC3 and TIF-2, were produced with the *E. coli* BL21(DE3)pLysS strain (Stratagene) containing the expression vector for each protein. The bacterial culture was stimulated with 0.25 mM isopropyl-β-D-thio-galactopyranosine for 3 h at 37ºC once it had reached an OD₆₀₀ of 0.4 to 0.6. The GST-proteins were purified and immobilized on glutathione-sepharose 4B beads (Amersham-Pharmacia) according to the manufacturer’s protocol. Elution was performed in the presence of glutathione and proteins concentrated using Millipore UFV5BGC00 centrifugation filter-tubes (Millipore Corporation, Espoo, Finland). The amount of GST-protein required for the assays was determined empirically.

4.2.1.4 Gelshift and supershift assays

For gelshift and supershift assays the following reaction was incubated for 15 min at RT:

5 µl *in vitro* translated proteins (10 ng)
2.5 µl KCl (1 M)
2 µl solvent or ligand
Materials and Methods

4 µl 5 x buffer (50 mM Hepes, pH 7.9; 5 mM DTT; 25% glycerol and 125 ng/µl herring sperm DNA (Promega))

5.5 µl H₂O

For supershift assays indicated amounts of the co-regulator GST fusion proteins were added to the reaction mixture and incubated an additional 15 min at RT.

After the incubation (both assays), 1 ng of [³²P]-labeled double-stranded oligonucleotides (50,000 cpm) were added and incubated 15 min at RT. Protein-DNA complexes were then resolved by electrophoresis through 8% non-denaturating polyacrylamide gels in 0.5 x TBE (45 mM Tris, 45 mM boric acid, 1 mM EDTA, pH 8.3)(200 V, 105 min). The gels were dried, exposed on a phosphoimager screen overnight and quantified on a Fuji FLA3000 reader using ImageGauge software.

4.2.2 Ex vivo methods

4.2.2.1 RNA extraction, cDNA synthesis and real-time PCR

Cells were seeded into 6-well plates and grown overnight to reach a density of 60-70%. The cells were stimulated with ligands as indicated in the individual publications. Total RNA was isolated from the cells using RNAeasy kit (ZymoResearch, HiSS Diagnostics GmbH, Freiburg, Germany) as instructed by the manufacturer. Total RNA amount was quantified (NanoDrop ND-1000, NanoDrop, USA) and 1 µg used in the following cDNA synthesis reaction:

1 µl oligoDT₁₈ primer (100 pmol)
4 µl dNTPs (5 mM)
2 µl DTT (10 mM)
8 µl 5 x M-MuLV buffer
1 µl RNasin (40 U/µl, Fermentas)
2 µl M-MuLV reverse transcriptase (20 U/µl, Fermentas)
ad H₂O to 40 µl

The reaction was incubated for 1 h at 37ºC followed by an inactivation for 5 min at 95ºC. The cDNA was diluted 1:10 in H₂O.

Real-time quantitative PCR was performed in an IQ-Cycler (BioRad) using the dye SybrGreen (Molecular Probes, Leiden, Netherlands). The following reaction was set up:

2 µl 10 x buffer (HotStart PCR buffer (Fermentas) used in II, III, FastTaq PCR buffer (Roche) used in IV)
2.4 µl 25 mM MgCl₂ (replaced with H₂O for FastTaq reaction)
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes

0.4 µl dNTPs (10 mM)
1.2 µl SybrGreen (1:2500 dilution from stock)
0.2 µl HotStart / FastTaq DNA polymerase (5 U/µl)
4.8 µl H₂O
4 µl template cDNA
5 µl primer mix (0.8 µM)

PCR cycling conditions and respective primers are described in the individual papers but for most PCRs the following program was used:

1. Denaturation for 5 min at 95ºC
2. PCR amplification repeated for 40-45 cycles
   - Denaturation for 30 s at 95ºC
   - Annealing for 30 s at primer-specific annealing temperature
   - Elongation for 40 s at 72ºC
3. Final elongation 10 min at 72ºC
4. Denaturation for 1 min at 95ºC
5. Melt curve analysis with 0.5ºC decrease in 70 temperature steps

Fold inductions were calculated using the formula $2^{\Delta\Delta C_t}$, where $\Delta\Delta C_t = \Delta C_t(\text{treatment}) - \Delta C_t(\text{solvent})$, $\Delta C_t = C_t(\text{gene}) - C_t(\text{ARP0})$ and $C_t$ is the cycle at which the threshold is crossed. PCR product quality was monitored using a post-PCR melt curve analysis at the end of amplification cycles.

4.2.2.2 Chromatin immunoprecipitation assay

Cells were seeded into culture bottles and grown overnight to reach a density of 60-70%. At the start of the experiment the medium in the bottles was reduced to 10 ml. The cells were stimulated with ligands as indicated in the individual publications (between 0.5 and 4 h).

Cross-linking of proteins and DNA

Formaldehyde was added to culture medium to cross-link proteins bound to DNA as follows:
270 µl formaldehyde (final concentration of 1%), incubating for 5 min at RT (except in II for 10 min)
1.5 µl glycine (1 M), incubating for 5 min incubation at RT (to stop the cross-linking)

Both incubations were performed on a rocking platform. The medium was removed and the cells washed twice with ice-cold PBS. Subsequently, the cells were scraped into ice-cold PBS and centrifuged for 4 min at 700 g at 4ºC.
Lysis of cells and sonication

After centrifugation the cell pellet was resuspended in 1 ml of lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl, pH 8.1) supplemented with a protease inhibitor cocktail (Roche). Lysis was performed for 15 min at RT. The lysates were sonicated to result in DNA fragments ranging from 200 to 1000 bp in length (Diagenode Bioruptor, Liege, Belgium). Cellular debris was removed by centrifugation for 10 min at 20,000 g at 4°C.

Immunocollection

The lysates were diluted 1:10. The following incubation was carried out in presence of specific antibodies:

- 100 µl of undiluted lysate
- 5 µl (200 µg/ml) of indicated antibody
- 24 µg of sonicated salmon sperm DNA
- 900 µl ChIP dilution buffer (0.01% SDS, 1.1% Triton-X 100, 1.2 mM EDTA, 16.1 mM NaCl, protease inhibitors, 50 mM Tris-HCl, pH 8.1)

overnight at 4°C with rotation

The antibodies used are described in the individual publications. For input samples, 25 µl of undiluted lysate was diluted with 475 µl ChIP dilution buffer and the processing of these samples was continued with reverse cross-linking and DNA extraction as described below.

Collection of immunocomplexes

The immune complexes were collected with 60 µl of Protein A agarose slurry (Upstate Biotechnology, Lake Placid, NY, USA) for 2 h at 4°C. The beads were pelleted by centrifugation for 1 min at 100 g at RT and washed for 5 min on a rotating platform with 1 ml the following buffers in the following order:

1. Low salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 150 mM NaCl, 20 mM Tris-HCl, pH 8.1)
2. High salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 500 mM NaCl, 20 mM Tris-HCl, pH 8.1)
3. LiCl wash buffer (0.25 mM LiCl, 1% Nonidet P-40, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris-HCl, pH 8.1).

Finally, the pellets were washed twice with TE-buffer (1 mM EDTA, 10 mM Tris-HCl, pH 8.1).

Elution and reversal of cross-linking

Elution of immunocomplexes was performed twice with 250 µl of elution buffer (1% SDS, 100 mM NaHCO₃) for 15 min at RT with rotation and the eluents were combined. Cross-linking was reversed
by an overnight incubation at 65°C. NaCl was added to a final concentration of 200 mM into the incubation reaction and the incubation was followed by another incubation for 1 h at 45°C with proteinase K (final concentration 40 μg/ml, Invitrogen) (II). Alternatively (used in III and IV) the reverse-crosslinking was done in one step with 2 μl of proteinase K (final concentration 80 μg/ml, Fermentas) that was active at 65°C.

**DNA extraction**

The DNA was extracted by adding 500 μl of 25/24/1 phenol/chloroform/isoamyl alcohol followed by a centrifugation at 20,000 g at RT. DNA was recovered from the aqueous phase using a DNA extraction kit (Qiagen II Gel Extraction, Qiagen) as instructed by the manufacturer in paper II, or with precipitation with 0.1 volumes of 3M sodium acetate, pH 5.2 and 2 volumes of ethanol using 1 μl of glycogen (20 mg/ml, Fermentas) as carrier (III, IV). The output samples were dissolved into 30 μl H2O and the input samples into 200 μl.

4.2.2.3 **PCR of chromatin templates**

PCR conditions for primers are described in the individual papers (II, III, IV). PCR reactions were performed as described earlier (see ‘‘RNA extraction and real time PCR’’) in an IQ-Cycler with SybrGreen dye and the products resolved on 2% TAE (200 mM Tris-HCl pH 7.5, 100 mM acetate, 5 mM EDTA) gels. The gels were imaged with Fuji FLA3000 reader using ImageGauge software. The melt curve analysis combined with gel imaging allowed the detection of specific primer-dimer-free PCR products that were used for calculating from Ct-values the quantitative data relative to controls (used in IV). Output samples were first normalized to their inputs and subsequently the fold change relative to non-specific IgG background was calculated. The fold inductions were calculated using $2^{\Delta\Delta Ct}$, where $\Delta Ct$ is $Ct$(specific antibody) - $Ct$(IgG control antibody) and $Ct$ is the cycle at which the threshold is crossed. Relative association levels were calculated using $2^{-(10-Ct(output-input))}$.

4.2.3 **In vivo methods**

Eight-week-old male BALB/c x DAB2 mice (National Laboratory Animal Center, Kuopio, Finland) were housed in stainless steel metabolic cages under controlled temperature (21–23°C) and light conditions (lights on 7 a.m. to 7 p.m.). Mice had free access to water and diet ad libitum (Altromin, Lage, Germany) for 14 days prior to initiation of treatment. All experiments were approved by the Committee for the Welfare of Laboratory Animals at the University of Kuopio and conducted in accordance with the guidelines of the European Community Council directives 86/609/EEC. GW501516 was administered in saline by intraperitoneal injection (1 μg/g body weight). After 3 and 6 h the animals were sacrificed and their livers were removed and shock frozen in liquid nitrogen.

Liver tissue samples were pre-homogenized in lysing matrix A tubes (Bio 101, Vista, CA) using a Fast Prep FP120 machine (Savant Instruments, Holbrook, NY, USA). Samples were processed twice for 40 s at setting 6.0 with a 10 min cooling interval on ice. Afterward, the samples were cooled on ice for 10 min, spun down for 1 min at 2000 rpm in a bench-top centrifuge and 600 μl of the cleared supernatant was transferred to the Mini RNA isolation kit (ZymoResearch) columns for RNA
4.2.4 In silico methods

4.2.4.1 Structural analysis and visualizations of the protein structures

For visualization and structural analysis of crystal structures from the PDB, the program SwissPDB viewer (www.expasy.org/spdbv/) was used. This software allows the visualization of secondary structure elements and inter/intramolecular interactions, such as hydrogen bonding.

4.2.4.2 In silico screening of putative PPREs

Genomic sequences spanning +/-10 kb around the TSS (III, IV), or until the end of the gene (II), of the genes analyzed were extracted from the current database release (NCBI build 35 (II), Ensembl release 40 (III, IV) from human genome (II, III, IV) and mouse genomes (IV) (for the genes mSULT2A1 and mRETN a full sequence became available in the October 2006 release 41 and was used for the analysis). Putative VDREs/PPREs were screened from the sequence files as described in the individual publications (II, III, IV). Conservation of putative REs between human and mouse was checked (III, IV) using the Vertebrate Multiz Alignment and Conservation track available from UCSC genome (NCBI releases: human Feb 2006 hg18, mouse Feb 2006, mm8) (Blanchette et al., 2004). The PPRE was marked conserved when a putative PPRE was predicted by the search in both species at an aligning location. Conservation of surrounding sequence was checked 50 bp upstream and downstream of the TSS: two occurrences of a continuous stretch of minimum five matching bp were required to label the PPRE to be located within conserved surrounding sequence. The human chr 19 and its syntenic mouse regions, used in IV, were extracted from Ensembl release 41 and screened for putative PPREs (medium to strong predicted binding strength) +/-10 kb around each TSS. The Ensembl ortholog prediction was used to match the respective human and mouse genes.

A list of established human PPAR target genes (used in IV) was extracted from literature with the following selection criteria: 1) mRNA or protein level changes in human were reported and 2) a PPRE was described from the human gene and 3) tested in at least gelshift, reporter gene or ChiP assay.

4.2.4.3 Calculation of predicted binding strength

Prediction of relative binding strength of PPREs (used in IV) was based on experimental gelshift results. Data of single nucleotide variations obtained in study III was used to classify them according to reduction in binding strength to three categories (IV, Table S1). Many natural sequences contain more than one deviation from consensus. To approximate the combination effect of variations from the different categories experiments were carried out with representatives from different combinations of these categories. An extended collection of binding strength data of PPRE-like sequences was used to calculate the average binding strength in each of the following combinations 1/0/0, 0/1/0, 2/0/0, 3/0/0, 1/1/0, 0/0/1, 0/2/0, 2/1/0, 3/1/0, 4/0/0 and 1/0/1 of variations from category I/II/III respectively. Other combinations resulted in less than 1% average binding and were not
considered for the PPRE search. Assignment of each tested sequence to these combination classes was made according to IV, Table S1, subtype specifically.

4.2.4.4 Clustering using self-organizing maps (SOMs)

SOM is an artificial neural network algorithm that belongs to the unsupervised learning methods. The algorithm produces a mapping of high-dimensional data into two-dimensions. Therefore it is useful in the visualization and interpretation of large data sets (Kohonen, 1997). The map consists of a regular grid of processing units, or “neurons”. These are initially assigned some arbitrary values in the data space and in the following iterative cycles move towards the actual data points. Three basic steps are repeated at the SOM algorithm until it converges: sampling, similarity matching and updating.

1. Initialization. Choose randomly the initial weight vectors \( w_j(0), j = 1, 2, \ldots, l \), of the l neurons in the lattice.
   - Alternatively, the weight vectors may be chosen randomly from the available input (data) vectors \( x_1, \ldots, x_N \).

2. Sampling. Take a sample vector \( x(n) \) from the input space for the iteration \( n \).

3. Similarity matching. Let \( i(x) \) denote the index of best matching (winning) neuron for the sample vector \( x \).
   - At iteration \( n \), \( i(x) \) is found from the minimum Euclidean distance criterion \( i(x) = \arg \min_{j = 1, 2, \ldots, l} \| x(n) - w_j \| \).

4. Updating. Update the weight vectors of all neurons using the rule \( w_j(n + 1) = w_j(n) + \eta(n) h_{ij}(x(n))[x(n) - w_j(n)] \),
   - Both the learning parameter \( \eta(n) \) and the neighborhood function \( h_{ij}(x(n)) \) are varied during learning.

5. Continuation. Continue with step 2 until the feature map has converged.

The SOM algorithm leads to a topological ordering of the feature map in the input space. Regions, in which the data are dense, occupy a larger domain in the output space.

A variation of the SOM, called tree-structured SOM, was used here. In this algorithm several SOMs are produced that are organized hierarchically in a pyramid-like fashion in several layers. The number of neurons at a larger level is four times the number of the previous level. The main difference in the algorithm is that the neighborhood function has been reduced to a fixed form, where always the four adjacent neurons of the best matching unit form the neighborhood that gets updated. The lower levels of the map give a coarse mapping of the data, with fine structure and clusters emerge as more neurons are included in the map.

4.2.4.5 Sammon’s mapping

The aim of Sammon’s mapping is similar to that of SOM, to represent points from an n-dimensional
Materials and Methods

space in 2-dimensions. Compared to SOM, some more accurate projection is produced where the conservation of the original structure is maximized. This quality means that the mapping is able to represent the relative distances of vectors and is useful in determining the shape of clusters and relative distances between them. The calculation of the map, however, is more time consuming. This can be solved by applying SOM before Sammon’s algorithm, where the SOM has achieved data reduction by replacing the original data vectors with fewer representative reference vectors.

A version of the Visual Data software (Visipoint, Kuopio, Finland), whose core is based a tree-structured SOM, was used (Koikkalainen, 1994). SOM was applied to cluster binding site composition data of target genes with the aim to separate the genes based on number of medium/strong binding sites and conservation pattern of all binding sites. This dataset consisted of six variables. The first two variables BS\(_H\) and BS\(_M\) represented the sum of predicted binding strength of \(n\) putative medium/strong PPREs found within the 20 kB of the analyzed region, in human and mouse respectively. The maximum of the subtype-specific predicted values was chosen for each PPRE. BS = bs\(_1\) + bs\(_2\) +... +bs\(_n\), where bs\(_i\) = max \{bs\(_{\alpha}\), bs\(_{\gamma}\), bs\(_{\delta}\)\}. The remaining four values represent conservation patterns. Binary values were used, where 0 denotes absence of conserved binding sites and 1 presence of them: CS\(_H\) = conserved medium/strong binding sites in human, CW\(_H\) = conserved weak binding sites in human, CS\(_M\) = conserved medium/strong binding sites in mouse, CW\(_M\) = conserved weak binding sites in mouse. Prior to SOM initialization, all variables were scaled between 0 and 1. The maximal resolution was set to 256. An initial coarse map at resolution of 4 was used in cluster assignment and assessment of their separation was done using Sammon’s mapping. The final clustering of the gene data was based on the 256-map and respective Sammons mapping, with the cluster numbering referring to original cluster at the low-resolution map.
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes
5. Results

5.1 Ligand-independent association of human PPARs with CoAs

The positioning of the last α-helix of NRs, H12, has a critical role in the association with CoAs and the subsequent transactivation events. Comparison of NR crystal structures shows that this part of the receptor undergoes largest conformation changes upon activation. In the case of many receptors, such changes happen upon ligand binding (as is evident from structural differences between unbound apo-structures and those crystallized with an activating ligand, or agonist). However, some NRs are characterized to have constitutive activity, which can be attributed to an ability to achieve the active conformation (and H12 positioning) without a ligand. Consequently, these receptors can associate with CoAs in a ligand-independent fashion.

PPAR ligands are known to directly contribute to the stabilization of the active conformation by a direct contact made between the molecules and the H12 of the receptor. However, a structural comparison of the positioning of H12 in the apo-structures of PPARγ and β/δ and NRs with known constitutive activity (LRH-1 and ERRγ) revealed a surprising similarity in the positioning of the helix (I, Fig. 1A). Furthermore, as the apo-PPARγ structure was superimposed with the respective agonist-bound structure, the critical separation of the charge clamp residues, K329 and E499, was preserved in the apo-structure (I, Fig. 1B and C). This structural feature suggested that the CoA interface is preserved in the unliganded receptor.

The closest NR relative to the PPARs, with characterized constitutive activity, is CAR. Unfortunately, at the time of this study, a CAR crystal structure was not available for direct structural comparisons. However, this receptor, whose constitutive activity we had addressed in earlier studies (Frank et al., 2004), was included into a comparative panel to assess the CoA association of PPARs, together with VDR, a NR that does not bind these proteins in the absence of ligand.

Supershift experiments with members of the p160 CoAs SRC-1, TIF-2 and RAC3, demonstrated that all PPARs were able to interact in absence of ligand (I, Fig. 2A), with RAC3 showing the most prominent interaction. Addition of ligand enhanced the interaction, however the effect was minor. The association of PPARγ was compared to that of CAR and VDR with increasing concentrations of GST-RAC3 protein (I, Fig. 2B). Without a ligand, CAR associated already with the lowest concentrations, whereas VDR did not associate at all. PPARγ had an intermediate profile that demonstrated its ability to interact with CoAs without a ligand. Addition of a ligand had no effect on CAR, a minor effect on PPARγ and a drastic effect on VDR that enabled its association, even with the lowest CoA concentrations tested. The stabilization of an active conformation of PPARs, ligand-independently, was also apparent from the limited protease digestion assays performed (I, Fig. 2D).
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing *in silico* methods into the analysis of nuclear receptor regulomes

Whereas the binding of a specific ligand was necessary for the stabilization of the LBD of the VDR (presence of c1-band), this was not needed for CAR and of limited effect concerning PPARγ.

Next, the basal activity and ligand-inducibility of the chosen NRs, were compared using cell-based assays. Two cell lines, the human breast cancer cell line MCF-7 and the human embryonic kidney cell line HEK293, were used. The cells were transfected with a luciferase reporter construct bearing four copies of the characterized DR1-type RE, derived from the human CPTIβ gene promoter (for PPAR binding), two copies of the DR4-type RE from rat pituitary specific 1, Pit-1, promoter (for CAR binding), or four copies of the DR3-type VDRE of the rat atrial natriuretic factor (ANF) and the respective expression constructs (I, Fig. 3A-D). High activity levels in absence of ligand were seen in both cell lines in case of both CAR and PPARs, but not in case of VDR. Addition of ligand showed a 36- and 74-fold increase in activity levels with VDR, CAR was stimulated maximally 1.8-fold and PPAR agonists reached a 3.5-fold maximum (PPARβ/δ agonist). To show that this effect was not due to presence of endogenous PPAR agonists in the cells, we overexpressed the CoR NCoR1. A significant decrease in activity levels was observed. However, this effect was reversed by the addition of strong agonists, such as rosiglitazone in the case of PPARγ.

In conclusion, ligand-independent association of PPARs with CoAs was shown *in vitro* with supershift and LPD assays and with transfection assays in living cells.

5.2 Structural determinants of H12 stabilization in PPARs

Structural *in silico* analysis of the apo- and agonist-bound PPAR structures was used to derive candidate amino acids responsible for the ligand-independent CoA association. Mutants were subsequently prepared to test their significance in transfection and supershift experiments (I, Figs. 4-6). Candidate amino acids were assigned to four groups, based on which type of stabilization they provide (PPARγ numbering is used for the residues). The two first groups contain amino acids identified to contribute to H12 stabilization (I, Fig. 4). The first group consists of K329 and E499 that stabilize the active conformation both ligand-dependently and -independently. Mutation of either of these amino acids abolishes the CoA interaction and the luciferase activity levels of the mutant receptors, inside cells, decreases dramatically. The residual activity level was comparable to that of a H12 deletion mutant (30% versus 25%). This remaining activity suggests that interactions with other types of CoAs that associate with other parts of the receptor, such as the AF-1 domain, are possible and contribute to the normal functioning of PPARγ. The second group consists of K347 and D503, which interact directly with each other and contact the CoA peptide. These residues were shown to affect the ligand-independent association of the NR with CoAs only.

The third group includes residues involved in the heterodimerization of PPARγ and RXR, in addition to their role in the direct stabilization of H12 (I, Fig. 5). The charged amino acids, E352, D424, R425 and R471, form salt bridge pairs that stabilize the tertiary structure and are strictly conserved in all heterodimerizing NRs. Y505 is in turn hydrogen bonded to these residues providing a structural link.
between the heterodimerization interface and H12 stabilization. The individual mutations of these residues reduced basal activities to between 45 and 80% of the wild type receptor value and are all ligand-responsive.

The residues in the last group, H351, Y355, H477 and Y501, contact the ligand (rosiglitazone) and additionally form an extensive hydrogen bond network in the absence of agonist (I, Fig. 6). In effect, they are responsible for both ligand-dependent and -independent stabilization. The mutants showed slightly lower basal activity (80%) and reduction in ligand-binding ability (45%). These mutants include the mutation of Y501, the residue responsible for the direct ligand-H12 interaction. However, much stronger effects were seen with its hydrogen-bonding partner H477.

A sequence comparison with PPARα and β/δ shows that all of these residues identified for the stabilization of H12 in PPARγ are conserved (I, Fig. 7). Further, phylogenetic trees based on sequence alignments of NR H12 residues (8 last residues) demonstrate that H12 of PPARs resembles most that of CAR (Fig. 8). In conclusion, the four groups of amino acids described are responsible for the ligand-independent stabilization of H12 in PPARs and the phylogeny supports the conclusion that the structural and functional features of the H12 of PPARs most resemble those of the constitutively active receptor CAR.

5.3 Identification of direct VDR and PPAR target genes within the IGFBP gene family

A number of NRs have been reported to regulate the IGFBP gene family members, suggesting a central role for this family in NR driven physiology. In total, this gene family has six members and of which two pairs, IGFBP-1 and IGFBP-3 (on human chr 7) and IGFBP-2 and IGFBP-5 (human chr 2), are direct genomic neighbors to each other, oriented in a tail-to-tail direction. Previous studies indicated that the IGFBP-3 gene is a direct VDR target gene and the protein product was shown to play a role in mediating the anti-proliferative effects of 1α,25(OH)2D3. Briefly, the physiological role of this gene family is in the regulation of the bioavailability of the circulating growth factors IGF-I and IGF-II and in the regulation of their effect on membrane signaling (Kelley et al., 1996). Additionally, the IGFBP-1 gene has been characterized as a marker for metabolic diseases, fitting with the PPAR regulatory domain (Gibson et al., 1996). Less evidence for the regulation of family members by PPARs is available, though, one research team has shown evidence for IGFBP-1 gene regulation by PPARγ ligands in the ovary (Seto-Young et al., 2005). However, this effect may be mediated by the NR PXR (Hilding et al., 2003). Since several IGFBP family members exist and multiple isotypes may be expressed in a given tissue, the overall physiological effect could reflect regulation of multiple members, rather than just one. Therefore, in the following studies we tested the responsiveness of all members of this family to VDR (II) and PPAR (III) ligands. Additionally, we expected that some members may be differentially responsive to these NRs, making the gene family an interesting test set to try to correlate the mRNA inducibility with binding site composition as revealed by in silico search of VDREs and PPREs.
5.3.1 Regulation of IGFBPs by VDR

The two cell lines used in this study (II) were the prostate cancer cell line PC-3 and the osteosarcoma cell line SaOS-2. The expression of all family members was confirmed by quantitative, real-time PCR from both cell lines (II, Fig. 1A and B) and the relative expression levels showed that IGFBP-4 and IGFBP-6 were the highest expressed. In PC-3 cells, IGFBP-2 and especially IGFBP-5, had relatively low expression, whilst IGFBP-1 and IGFBP-3 had moderate expression levels. While all members were also expressed in SaOS-2, the differences were less pronounced. To exclude possible secondary effects arising from the induction of these genes on account of primary 1α,25(OH)2D3 target gene stimulation, the inductions of the family members were tested both in the presence and absence of the protein translation inhibitor cycloheximide at 24 h (II, Fig. 1C and D). The induction pattern was compared to that of p27, which is known not to be a gene that is directly regulated by VDR. As seen from both cell lines, IGFBP-1, IGFBP-3 and IGFBP-5 responded in the presence and absence of cycloheximide, while inductions of IGFBP-4 and IGFBP-6 (only observed in SaOS-2) were abolished by the cycloheximide treatment. Additionally, IGFBP-1, IGFBP-3 and IGFBP-5 were the only family members that were inducible with shorter treatments, a typical characteristic of direct target genes. In conclusion, we identified three responsive family members that are likely direct targets and three non-responsive members.

In order to search for possible VDREs in the regulatory regions of these IGFBP genes, we analyzed a list of known natural VDREs (>15) and obtained on that basis the consensus sequence RGDKYR (R = G or A, D = A, G or T, K = G or T, Y = C or T) which we searched for REs in DR3-, DR4- and ER6- to ER9-orientations (orientations were based on our previous in vitro studies) from the IGFBP gene areas (II, Fig. 2). We included 10 kB of promoter sequence and the coding sequence (until end of last exon). Additionally, for the gene pairs IGFBP-1 and IGFBP-3 and IGFBP-2 and IGFBP-5, we analyzed their intergenic region. Moreover, only candidate VDREs, formed of hexamers with maximal two deviations from the optimal RGKTCA core sequence, were included. Additionally putative VDREs located in Alu repetitive sequence elements were excluded due to the high likelihood that they mostly lie within closed chromatin are inactive (Kondo and Issa, 2003). With these filtering conditions, we found several putative VDREs from the promoters of the genes IGFBP-1, IGFBP-3 and IGFBP-5. The IGFBP-5 gene also had an element in its first intron, in addition two putative VDREs were located further in the first intron of IGFBP-2. Furthermore, both intergenic regions analyzed contained two putative VDREs. In contrast, IGFBP-4 and IGFBP-6 contained no VDREs that passed our criteria.

Next, we tested whether the putative VDREs (15 in total, out of which one contained a SNP that resulted in two alternative forms of that VDRE) were able to bind the receptor in transactivation assays. The MCF-7 cell line was used for comparison together with the cell lines PC-3 and SaOS-2. Each candidate VDRE was cloned in two copies into a luciferase reporter construct and the inductions compared to that of the known rat ANF DR3-type VDRE (II, Fig. 3). Overall, higher inductions were obtained in the MCF-7 cell line, otherwise similar effects were observed: REs 2, 8 and 13 (for numbering refer to II, Fig. 2) were comparably or higher induced, relative to the reference
Results

VDRE, the SNP variation of RE13 lost responsiveness compared to the more common variant but yet was functional and additionally REs 3, 4, 5, 9, 10 and 14 were induced 5- to 13-fold indicating that they could also function as VDREs. These VDREs are located in the vicinity of the responsive IGFBP family members (II, Fig. 2). However, both IGFBP-2 intronic elements were shown to be able to function as VDREs as well. REs 3, 8 and 9 are DR3-type REs, where as REs 10 and 14 are DR4-type REs (II, Table 1). Importantly, these results show that ER7- (REs 2, 4, 5) and ER9-type (RE13) REs can function as VDREs.

The binding of VDR-RXR heterodimers to these putative VDRE sequences was also assayed with gelshifts (Fig. 4). Again the rat ANF gene DR3-type VDRE served as a reference. The results were in accordance with the transactivation assays concerning REs 2, 4, 5, 8, 9, 10 and 13 (both variations). Of these, REs 2, 4, 5, 8, 9 and 13 bound stronger than the reference RE. The ER11-type RE created from RE13 by the SNP bound 44% relative to the ER9-type variant. Together with the transactivation results, this suggests that VDR-RXR indeed has the capacity to bind a large variety of ER-type elements. Two REs (3 and 14) that were functional in the transactivation assay bound VDR-RXR 24% and 11%, respectively, and would therefore be considered too weak to be efficient VDREs based on earlier in vitro criteria (Toell et al., 2000). In summary, our in silico search identified a number of strong VDREs and overall a good correlation with both methods was obtained concerning functionality of the predicted VDREs. Only in case of two REs (11,15), their functionality was only evident from one assay (gelshift).

To assess the functionality of these elements in authentic chromatin contexts, ChIP assays were performed (II, Figs. 5A-C). Chromatin was extracted from PC-3 cells treated for 0.5, 1, 2, 3 and 4 h with 1α,25(OH)2D3. Input lanes serve as control of similar detection capacity of the primers and IgG and p53 antibodies were used as unspecific controls. Even though the in vitro assays already suggested that multiple VDREs had been identified that could serve as functional binding sites, surprisingly many regions associated with VDR in the ChIP assay as well and only REs 11, 12 and 15 failed to do so. Several regions (2, 3, 6/7, 8, 9, 10, 13 and 14) bound the receptor in the absence of a ligand. Modulation in the binding was seen for regions 2, 8, 9 and 10, with association reaching a maximum at 120 min and stronger yet for regions 1, 4 and 5. Further ChIP assays were carried out to assess the occupancy of these sites by other regulatory proteins. The binding of RXRα was seen in all regions (apart from RE1) that associated with VDR. The CoR SMRT was detected at 8 of 10 regions that associated with VDR-RXR in the absence of ligand and reappeared at regions 4, 5, 13 and 14 at later timepoints. CoA binding (SRC-1) was seen 0.5-2 h after ligand treatment. However, both co-regulators had an individual profile for their association to different regions. In contrast, the association of the mediator protein and p-PolII was timely coordinated for all regions, occurring 1 to 2 h after ligand treatment and persisting for approximately 1 h. Taken together, the IGFBP-1/IGFBP-3 and IGFBP-2/IGFBP-5 gene clusters contain 10 VDR- and RXR-associated regions (2, 3, 4, 5, 6/7, 8, 9, 10, 13 and 14), which each show an individual, ligand-dependent profile of SMRT, SRC-1, TRAP220 and p-PolII binding.
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes

Since the two REs located in the intron of IGFBP-2 showed functionality in all of the assays described above, this raised the question, whether they are involved in the regulation of the IGFBP-2 gene that is greater than 17 kB away, or may regulate the even more distal IGFBP-5. The answer to this question was provided indirectly by assessing with ChIP assays the VDR association with the TSS regions (II, Fig. 5C). VDR interaction was found on the TSS of IGFBP-1, IGFBP-3 and IGFBP-5 and not on the TSS of IGFBP-2, which confirmed the real-time PCR data that showed that these three genes and not IGFBP-2, are VDR regulated.

In conclusion, the in silico analysis performed well to detect putative VDREs and was in good correlation with the real-time PCR analysis assessing the regulation of mRNA levels of the IGFBP family members.

5.3.2 Regulation of the IGFBPs by PPARs

None of the IGFBPs had been characterized as direct PPAR targets at the time of the study. Therefore, we selected as a starting point the IGFBP-1 gene for further analysis, for which already some evidence for regulation existed. Using HepG2 cells treated with specific agonists for the three subtypes, we showed that the IGFBP-1 mRNA level was induced 2- to 3-fold in an early time-course experiment. The effect was also seen in mouse liver tissue obtained from mice treated with the PPARβ/δ agonist GW501516 (III, Fig. 1). The induction levels were comparable with the known PPAR target gene CPT1α.

For the purpose of developing an efficient binding site search for PPREs, we evaluated the binding of the three subtypes to single nucleotide variations of the consensus PPRE, AGGTCAaAGGTCA (III, Fig. S1). This resulted in a categorization of the binding variations into three classes I (able to bind 75 +/- 15% of consensus RE), II (45 +/- 15%) and III (15 +/- 15%). We used this classification to identify tolerated combinations of variations, using gelshift data on 10 published PPREs (III, Table S3) and an additional set of PPRE-like sequences (data not shown). Based on this data, we formulated an in silico search, where all putative PPREs that deviated maximally by three variations of class I combined with one from class II (if these were on the same site and the other hexamer fitted the consensus) or less variations (only one variation from class III was allowed and maximally two from class II) were identified.

A 20 kB region, centered around the TSS of the human IGFBP-1 gene, was analyzed in silico, which resulted in 5 putative PPREs (III, Figs. 2A). The binding of PPARs to these REs was examined with gelshift assays (III, Figs. 2B-D) and this revealed that most of the sites identified were fairly weak PPREs (binding 1 to 17% relative to the CPT1 PPRE used as reference), as was expected from the number and variation class of consensus variations in their sequences (III, Table 1). In general, the binding of PPARγ was stronger to each element than that of PPARα or PPARβ/δ.

Based on the gelshift data alone it seemed that only very weak associations with PPARs would be mediated by the PPREs identified. Next, the HepG2 cell line was used to test, whether this held true also for the associations in ChIP (III, Fig.3). An initial panel showing association with all PPARs (III,
Fig.3A) revealed that only the two promoter elements (RE1, RE2) associated with the receptor ex vivo. The association on RE2 was detected ligand-independently, whereas PPARs appeared on RE1 only after respective ligand stimulations (2 h). The specific ligands seemed to favor the association of their respective subtype on RE2 (except with no change in the PPARβ/δ association level was seen with PPARα agonist). In contrast, only a few bands in RE3 and 4 were detected, these are likely attributable to non-specific background. No signal was detected from the strongest in vitro PPRE containing region 5, or the control region that contained no PPRE. In addition, the association of other regulatory proteins was evaluated with a longer time course of PPARα treatment. RE1 was shown to associate with a restricted panel of regulatory proteins (PPARα, RXRα and p-PolII), whereas RE2 associated with the full set of proteins tested (TRAP220, PGC-1α, RAC3 and NCoR1, in addition to the ones mentioned). The association of RXRα, p-PolII, TRAP220 and both CoAs were strongest at 120 min, in contrast to the diminished NCoR1 signal at this time point. At time point 0 min and at some later time points, simultaneous occupancy by both CoAs and CoRs was observed. However, based on the co-factor code such a co-existence is possible (Rosenfeld et al., 2006). Additionally, the results indicate that ligand-independent CoA association occurs in living cells. The structural explanation for this effect was explored in study I.

The regions that were found associating with PPARs in ChIP (around RE1 and 2) were cloned into luciferase expression vectors that were co-transfected with and without the corresponding PPAR isotype and RXRα into HepG2 cells and compared with PPRE containing region of the human CPT1 gene promoter (III, Fig. 4). Both the IGFBP-1 and the CPT1 promoter region constructs demonstrated a higher basal activity relative to the empty vector already in the absence of overexpressed NRs and this activity was further increased with the introduction of excess PPAR and RXRα. Furthermore, mutations of the critical 6th position of the PPAR binding hexamer (changing A to G) in the promoter region containing constructs, resulted in a significant loss of activity (both for RE1 and RE2). The mutation into the CPT1 construct was introduced at a less critical position (7th position, A to T) and had a minor effect on the resulting luciferase activity. Significant ligand-responses were detected with all PPARs on RE2, whereas only the PPARβ/δ ligand induced RE1. In conclusion, RE1 and RE2 both contribute to the activation of the IGFBP-1 promoter, with RE2 having a more prominent role.

The successful identification of functional PPREs from the IGFBP-1 gene prompted us to screen the regulatory regions of the five other human IGFBP gene family members in a similar fashion. Five putative PPREs were identified from the promoter of the gene IGFBP-2, two from the IGFBP-3 gene, one from the IGFBP-4 gene, three from the IGFBP-5 gene and seven from the IGFBP-6 gene. Based on the in silico result, IGFBP-2, IGFBP-5 and IGFBP-6 seemed therefore likely to be regulated, whereas IGFBP-3 and IGFBP-4 did not. This prediction was tested by monitoring the PPAR occupancy of the TSS regions of these five genes (and of IGFBP-1 as reference) with ChIP ex vivo (III, Fig. 5). Template from HepG2 cells was used, except in case of IGFBP-5 that was not expressed in HepG2 cells. Instead, we used the HEK293 kidney cell line to derive the chromatin immunoprecipitates. Additionally, the HEK293 template served as a control for IGFBP-1 TSS, since
the IGFBP-1 gene was not expressed in HEK293. The result confirmed the prediction that PPARs are associated with the promoters of the human IGFBP-1, IGFBP-2, IGFBP-3 and IGFBP-6 genes. PPARα was associated both in absence and presence of its ligand, whereas modulation was in most cases seen for PPARγ and PPARβ/δ association. PPARβ/δ recruitment was most dependent on its ligand and in the case, of the PPARγ recruitment to the human IGFBP-6 gene promoter, ligand addition abolished the association with the TSS region. None of the PPAR isotypes were detected on the IGFBP-3 and IGFBP-4 TSS containing regions. Similarly, neither on the IGFBP-1 gene TSS in HEK293 cells, nor on the IGFBP-5 gene TSS PPAR binding was detectable.

The success of the in silico prediction was also verified by real-time PCR in HEK293 and HepG2 cell lines (III, Fig. 6). In HepG2 cells, IGFBP-2 gene expression was induced at 4 h with all PPAR isotype selective ligands, whereas IGFBP-6 gene expression was significantly down-regulated at the 2 h time point with PPARβ/δ ligand (GW501516) stimulation. Very similar effects were seen in HEK293: IGFBP-2 induction reached significance at 4 h and in this cell line all PPARs down-regulated IGFBP-6 gene expression levels, though at different time points. IGFBP-5, the gene that was not expressed in HepG2 cells, was regulated significantly, already at 2 h. In neither of the cell lines, significant regulation of IGFBP-3 and IGFBP-4 gene expression was detected.

In summary, these two studies (II and III) identified and characterized three direct VDR targets and four PPAR targets, of which direct PPAR-mediated regulation was characterized in detail for IGFBP-1, amongst the IGFBP gene family. Of these, IGFBP-1 and IGFBP-5 are shared targets of VDR and PPAR activation. Moreover, the results indicate a clear correlation with in silico VDRE/PPRE predictions and regulations of mRNA levels by the respective receptors.

### 5.4 Meta-analysis of the binding site composition of human PPAR target genes

The basis of efficient in silico predictions of binding sites and target genes is a detailed understanding of target gene properties on sequence level. The characterization of PPREs from regulated gene promoters has resulted in a large collection of PPREs that deviate significantly from the consensus sequence. The ubiquity of such PPRE-like sequences on a whole genome level is in contrast to the number potential PPAR target genes in a physiological context (a few thousand per tissue) and the number of receptor molecules (estimates varying from few thousand per cell to values below hundred thousand). A recent effort to better model the binding preferences of PPARs used position weight matrices to describe all published PPREs (Lemay and Hwang, 2006). However, such an approach has limited ability to predict bona fide PPAR binding. In addition to binding strength, a number of additional parameters can be named that could influence binding site functionality. One common trend in the location of binding sites is a positional bias towards the TSS. This would be apparent from the collection of identified PPREs, but as discussed before is in contrast with a multi-genome comparison of NR binding site distribution (Xie et al., 2005). Further, a common approach for the detection of functional binding sites is to rely on conservation. However, the maintenance of
responsiveness may not require conservation of exact binding site composition as exemplified before with the Drosophila eve enhancer. The evolutionary models governing the preservation or turnover of binding sites have not been studied in detail, however as limits they will influence the design of in silico search methods. Therefore, in this study we aimed to explore the binding site composition of PPAR target genes and evaluate the contribution of each parameter mentioned, namely binding strength, location bias and conservation models.

At first we addressed the problem of relating binding site predictions with binding strength. This is not achieved in weight matrices and also fails when the number of variations increases with the affinity matrix approach. Therefore, we created based on experimental data a classification scheme that can approximate the binding resulting from the combinations of several variations. Each single variation from the consensus sequence had been evaluated earlier (III, Fig. S1) and resulted in a classification of variations to three classes (IV, Table S1) subtype specifically. A total of 116 sequences containing different variations from the AGGTCAaAGGTCA consensus sequence were classified according to Table S1 to the variation categories (number of variations from Class I, number of variations from Class II and number of variations from Class III) (IV, Table S2). The binding of PPARs to these sequences was tested with gelshift. The combinations 1/0/0, 2/0/0 and 0/1/0 are strong PPREs with 68, 40 and 44% binding respectively. Medium strength PPREs contain 3/0/0 or 1/1/0 variations resulting in 31 and 25% of binding. In addition, the combinations of 0/0/1, 0/2/0, 1/2/0, 1/0/1, 3/1/0 and 4/0/0 still resulted in observed binding (on average above 1%). Other combinations resulted in less than 1% average binding and were not considered. The standard deviations within the categories were between 1 and 15%, which suggest that good accuracy level can be reached with the prediction scheme.

We chose to address the contribution of binding strength, location and conservation with eight known PPAR target genes in detail. The selected genes contain the known up-regulated targets ACOX1 (Varanasi et al., 1996), CPT1B (Mascaro et al., 1998) and PPARα (Pineda et al., 2002). The downregulated APOC3 (Hertz et al., 1995) was included for comparison and in addition genes that contain unusual location or type of PPRE were selected: sulfotransferase 2A1 (SULT2A1) has a distal PPRE (Fang et al., 2005), ANGPTL4 PPRE is inside an intron (Kersten et al., 2000) and RVTrα PPRE is a DR2-type element (Fontaine et al., 2003). In addition, we included one regulated gene for which no PPRE has been described, UCP3 (Son et al., 2001), in order to test whether we could predict and verify the regulatory regions of this gene. The regulation of these genes was confirmed in our cellular models HEK293 and HepG2 (IV, Fig. S1). Each responded in both cell lines (APOC3 was not expressed in HEK293).

These eight genes were analyzed in silico for PPREs within 10 kB from their TSS (IV, Fig. 1). All PPRE categories that included PPREs with 5% or more binding for each subtype are shown. Those categories resulting in 1-5% of binding (1/0/1, 3/1/0, 4/0/0) are only indicated if the PPREs were conserved in mouse. For each PPRE the height of the bars indicates predicted binding strength, as determined based on the experimental binding data set. In total 46 PPREs were found with roughly an equal distribution relative to the TSS, ten PPREs were conserved in mouse and further six were inside
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes

conserved area (evaluated based on the surrounding sequence 50 bp left and right of the PPRE). The published DR1-type PPREs were successfully identified from the other genes, except ACOX1. However, the non-functionality of this ACOX PPRE has been reported (Woodyatt et al., 1999). In that study it was claimed that the human ACOX1 gene may not be at all a PPAR target, whereas here we show that the gene is regulated by PPARs and suggest five alternative binding sites, of which one is located in an evolutionary conserved area of intron 1. The in silico prediction of binding strength was confirmed for the UCP3 PPREs. This resulted in 16 out of 18 values to match the prediction with a deviation of less than 15%. The predictions of PPARγ and PPARβ/δ were all accurate while the two deviating values were underestimated for PPARα.

To evaluate further the functionality parameters we chose PPRE containing regions from these genes based on their location. Ten PPREs close to TSS (within one kB), ten further upstream and ten further downstream were selected. This resulted in 23 genomic regions that were cloned to luciferase reporter constructs (indicated in IV, Fig.1). The activity of the constructs in absence or presence of PPAR subtype expression vectors in response to PPAR subtype-specific ligands was tested by reporter gene assays in HEK293 and HepG2 cells (IV, Fig. 2). With the exception of the TSS of the RVRα gene, containing the DR2-type PPRE, eight of the proximal regions showed in at least one of the two cell lines significant inducibility by PPAR ligands. The region of the CPT1B gene region was inducible by all three PPAR subtype-specific ligands in both cell lines, whereas the seven other regions show a PPAR subtype- and cell type-specific profile. An increase of basal activity compared to empty cloning vector and its subsequent loss due to PPAR overexpression was observed with the proximal regions of the genes APOC3 and UPC3 in both cell lines and in addition in HepG2 cells with the intron 1 region of the ACOX1 gene and the proximal region of the APOA1 gene. A similar observation was made with the distal regions of the genes ANGPTL4, APOC3 and PPARα. From the upstream regions the region of the SULT2A1 gene was most active. In both cell lines the overexpression of PPARs clearly increased its basal activity and in addition a significant inducibility by all three PPAR ligands was observed. A similar observation was made in HepG2 cells for the distal region of the UCP3 gene, whereas in HEK293 cells response of this region was far more modest. In contrast, neither the distal regions of the genes ACOX1 and RVRα nor the region containing the published PPRE of the ACOX1 gene displayed any inducibility by PPAR ligands in any of the two cell lines. In addition, the distal regions of the genes ANGPTL4 and APOC3 were only inducible in HEK293 cells, whereas the PPARα gene responded only in HepG2 cells to GW501516 treatment. Within the five downstream regions the intron 2 region of the ACOX1 gene and the cluster region of the ANGPTL4 gene (containing four putative PPREs) showed in both cell lines a clear response to all three PPAR ligands, whereas the inducibility of the intronic region of the APOC3 gene was far more modest. Individual mutagenesis of the ANGPTL4 REs was carried out which demonstrated that the other REs in addition to the published RE contribute to the responsiveness of this region (data not shown). Finally, the cluster and intronic region of the UCP3 gene responded only in HEK293 cells to GW501516 treatment.
Results

The same 23 genomic regions were investigated by ChIP assays with chromatin extracts from HEK293 cells (or from HepG2 cells for regions from the APOC3 gene) that had been treated with solvent or for 120 min with the PPARα ligand GW7647 (IV, Fig. 3). We assessed these regions for the binding of PPARα, its partner receptor RXRα and p-PolII, i.e. for a sign for a direct link to the TSS. Chromatin templates were analyzed by quantitative real-time PCR and the specificity of the antibodies for the three proteins was compared with the non-specific background binding to IgG.

From the 23 tested regions the region of the CPT1B gene, the distal and published region of the ACOX1 gene, the distal 1, distal 2 and intronic region of the APOC3 gene and the cluster of the UCP3 gene did not show a specific binding of any of the three proteins. For the two regions of the ACOX1 gene this result confirmed their failure in the functionality test (IV, Fig. 2). The 16 other regions showed a significant association with PPARα in the presence of ligand. However, no statistically significant binding of p-PolII, neither in absence nor presence of ligand, was found in the published region of the APOC3 gene and in the distal regions of the genes ANGPTL4 and RVRα. When comparing the relative association levels of PPARα under this condition, we found most prominent binding to the region of the SULT2A1 gene, followed by the regions of the RVRα TSS and the proximal region of the PPARα gene (Fig. 3D). The latter two regions as well as the proximal regions of the genes APOA1 and UCP3, the distal region of the RVRα gene and the distal and intronic region of the UCP3 gene displayed a ligand-independent binding of PPARα.

In summary, from the 23 investigated genomic regions 17 show in HEK293 cells and 14 in HepG2 cells a significant inducibility by PPAR ligands. Twelve of these regions also associate in ChIP with PPARα, RXRα and p-PolII. Surprisingly, the most responsive regions are located either in far upstream or downstream areas of the genes. These results indicate that PPREs do not exhibit a strong positional bias towards the TSS. In addition, no preference of evolutionary conserved PPREs was found, but each of the genes contains at least one medium or strong PPRE. Three of the 12 fully functional regions, the ANGPTL4 cluster region, the proximal region of the PPARα gene and the distal region of the SULT2A1 gene, were already known before, but for the genes ACOX1, APOC3, RVRα and PPARα alternative/additional regulatory regions were identified. Finally, for the the UCP3 gene, for which so far no regulatory region was know, we identified three functional regions.

To extend the conclusions based on these eight genes, we extracted further 30 genes from literature that are known to respond to PPARs in humans (the criteria for selection required that regulation was shown in mRNA or protein level in humans and the human PPRE was evaluated in gelshift, reporter gene assay, or ChIP experiments). Each gene was analyzed in silico the same way as before. The orthologous mouse gene was included for comparison in order to identify traces of binding site evolutionary patterns. From these gene overview figures the location of the PPREs (x-axis) and their respective predicted binding strength (y-axis) is immediately obvious. In order to reveal further characteristics of the genes, such as overall similarity of their PPRE pattern and its evolutionary conservation, we clustered the genes with the SOM algorithm followed by Sammon’s mapping to identify clusters. The input dataset of the SOM consisted of six variables, which are the sum of the predicted binding strength (BS), the number of conserved strong and medium binding sites (CS) and
the number of weak binding sites (CW) both for the human and the mouse gene (IV, Table S3). An initial coarse map resulted in four clusters, which were each separated in the final map in up to five subclusters (Fig. 4A).

In general, clusters I to II contain genes that are well conserved between human and mouse, while the genes in cluster III are more divergent and those in cluster IV show no evolutionary conservation of PPREs. Cluster IA is formed of the genes ANGPTL4, APOA1, APOC3, CPT1A and phosphoenolpyruvate carboxykinase 1 (PEPCK1) (Duplus et al., 2003) that have both in human and in mouse multiple conserved regions with strong REs. A subset of genes further expanded this set: human CPT1A and mouse APOA1 show significant enrichment of de novo binding sites compared to their ortholog. In cluster IB, being represented by the genes LPL (Schoonjans et al., 1996) and LXRα, one RE region is conserved, but only the mouse gene contains further REs. UCP3 is found in cluster IC together with glycerol kinase (GK) (Patsouris et al., 2004; Lehrke et al., 2005), these both have multiple conserved RE regions. The distal UCP3 PPRE is conserved to a strong PPRE in mouse, but this element is out of the 10 kB window. On the other hand, the cluster of PPREs in human GK promoter seems to have lost significance in the mouse. The lipoprotein receptor-related protein 1 (LRP1) (Gauthier et al., 2003) gene represents cluster ID, in which multiple conserved REs are strong in mouse but weak in human. Together with the genes GK and LRP1, the genes found in cluster IE, caveolin 1 (Llaverias et al., 2004) and IGFBP1 (III), exhibit a retain-loss pattern concerning conserved PPREs, where only one species retained a cluster of strong PPREs. In case of LRP1, this appears to have arisen in the context of several compensating de novo binding sites. In contrast, cluster II contains genes with relatively fixed PPRE pattern: only one strong or medium conserved RE in human, which is conserved in comparable strength and location in mouse. This cluster is not subdivided into subclusters and contains the genes ADRP, CPT1B, CPT2, CYP1A1 (Sérée et al., 2004), G0S2 (Zandbergen et al., 2005), HMGCS2 (Iida et al., 2002), SR-B1 and spermidine/spermine N1-acetyltransferase (SSAT) (Ignatenko et al., 2004). The genes ADRP and CPT2 contain only the preserved strong RE. In addition to the conserved promoter cluster, the human CYP1A1 gene contains a stronger intronic PPRE, while its orthologs according to both database annotations have expanded in number of medium PPREs. Expansion of PPRE set only in one species is observed for mouse G0S2 and SSAT.

Cluster IIIA, which contains the genes cyclooxygenase 2 (COX2) (Pang et al., 2003) and semaphorin 6B (Collet et al., 2004), extends the pattern observed above (clusters ID and IE) with an opposite trend: the human ortholog contains one or two medium/strong REs, which are conserved but only weak in the mouse. Possible compensating element appeared only in the mouse ortholog of the latter gene. Also cluster IIIB contains one or two conserved REs, but they are weak both in human and in mouse. This cluster comprises the genes APOA2, APOA5, fatty acid desaturase 2 (FADS2) (Tang et al., 2003), PXR (Aouabdi et al., 2006), RVRα and SLC10A2. Interestingly, these genes each have novel binding sites at rather similar locations. The species-specific assembly of PPREs is most evident in cluster IV, where none of the PPREs are conserved. The ACOX1 gene represents cluster IVA, in which multiple strong, but non-conserved REs are found in both species. The genes resistin
(Patel et al., 2003) and SULT2A1 form cluster IVB, these have multiple REs in mouse and strong REs in human. The genes APOE (Galetto et al., 2001) and PPARα are in cluster IVC, which is characterized by one strong RE in the mouse ortholog and one or more non-conserved REs in the human gene. The remaining genes in IVD and E, CYP27A1 (Szanto et al., 2004), glutathione S-transferase A2 (GSTA2) (Park et al., 2004), transferrin (Hertz et al., 1996) and UDP-glycosyltransferase 1A9 (UGT1A9) (Barbier et al., 2003), carry one or two medium, non-conserved REs in both species and the p16 gene (CDKN2A) strong or medium non-conserved REs are found in mouse and but only weak REs in human. Taken together, although for some genes a conservation of PPRE pattern is evident, also a significant diversification of PPRE composition is visible.

In order to explore the evolutionary preservation patterns of PPREs further, the genes ACOX1 and ANGPTL4 were analyzed also from the genomes of chicken, chimpanzee, dog, rat and zebrafish (IV, Fig. 4B). The closest to the human genome is chimpanzee (Pan troglodytes), where four conserved PPREs are located in the ACOX1 gene, however the functional intron 2 region is missing. This human element was not conserved in any of the species analyzed, suggesting a novel human-specific PPRE. Mouse and rat (Rattus norvegicus) share two PPREs, however both also contain a unique set of further elements. The analysis of the other species is in accordance with the overall pattern of relatively species-specific composition of PPREs. Within mammals, in the ANGPTL4 gene the cluster of intronic REs is rather well conserved. The closest PPRE pattern in comparison to the human gene is observed in the dog (Canis familiaris). Two intronic PPREs are also present in rat, but a significant expansion in PPREs seems to have occurred in this species, including a distal consensus RE. The zebrafish ANGPTL4 gene is also profoundly enriched with strong PPREs, whereas the chicken gene has a quite poor PPRE content. Interestingly, the loss of the published PPRE is observed in the chimpanzee for this gene as well. Therefore, while this gene is an example of a much more preserved PPRE pattern, significant diversification is evident in some genomes and not all functional PPREs are conserved.

In conclusion, the SOM analysis and further genome comparisons together suggest that the integration of the stabilizing selection model into target gene characterization may more faithfully identify PPAR targets. Other common features of these genes include presence of strong PPREs or multiple (more than two) medium REs (74% of human genes, 68% of mouse genes). In addition, even though location of PPREs clearly does not determine functionality a slight bias to the immediate upstream region (500 bp upstream) could be detected (24% of genes).

In order to evaluate these parameters for the prediction of target genes, the most gene dense human chromosome, chr 19 (1445 known genes in Ensembl), was screened for medium and strong PPREs (based on PPARγ prediction). In parallel, the syntenic regions in mouse genome were analyzed (956 ortholog pairs). We identified 109 genes that contain in both species either a strong PPRE, multiple medium PPREs, or a proximal medium PPRE (within 500 bp upstream) (IV, Table 3). This approach has the ability to detect targets with divergent PPRE composition, at the same time the chance event that both species would contain random PPRE occurrences is low. An additional 71 genes were predicted using only one species, human (IV, Table S3). This list was limited to genes that show both
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes

enrichment (more than two PPREs) and contain at least one strong PPRE (such genes found from both species are in bold in Table 3).

From these lists the six human genes CYP4F8, LASS1, chicken ovalbumin upstream promoter transcription factor (COUP-TF) γ, p21-activated kinase 4 (PAK4), SLC27A5 and translocase of inner mitochondrial membrane 13 (TIMM13) were selected for the real-time PCR evaluation of their response to the PPARα ligand GW7647 in HepG2 cells (IV, Fig. 5A). After a 2 h ligand treatment all six genes showed a significant 1.8- to 4.2-fold upregulation of their mRNA. For a more detailed analysis we selected the LASS1 gene. The in silico analysis of the gene suggested four non-conserved PPREs, of which the two strong REs are in close vicinity (region 2) are the best candidates forming the PPAR-responsive region of the LASS1 gene (IV, Fig. 5B). Functional analysis of three genomic regions in reporter gene assays and ChIP in HepG2 cells indicated for region 2 a significant upregulation by PPARγ and PPARβ/δ ligands and an even more prominent basal activity of PPARα (IV, Fig. 5C), confirmed by ChIP (IV, Fig. 5D). A treatment with GW7647 induced significant binding of PPARα, RXRα and pPol II to region 2, but not to regions 1 and 3. This suggests that the two strong PPREs in region 2 mediate the response of the LASS1 gene to PPAR ligands.

In summary, the ability to predict binding site strength revealed a clearer picture of the PPRE composition of target genes extended over 20 kB of regulatory region. There is relatively weak preference for conserved binding sites, or proximal location of PPREs. Instead enrichment of medium and strong REs can capture a large majority of these genes. These features when exploited on a genomic scale led to the identification of six novel targets in human and the successful identification of functional PPREs of the UCP3 and LASS1 genes (in addition to further novel REs shown in the panel of seven known targets).
6. Discussion

6.1 Ligand-independent CoA association of the PPARs

The molecular mechanisms of NR activation were originally described for the group of NRs that are highly ligand-sensitive. These NRs are unable to interact with CoAs in the absence of a ligand. This ligand-dependence was attributed to the exact positioning of the C-terminal H12. Based on apo and ligand-bound structures of RARα and RARγ, respectively, an activation model was proposed where ligand-binding stabilizes the mobile H12, which then acts as a lid to the LBD (known as the “mouse-trap model”) (Moras and Gronemeyer, 1998). Later on, as more NR structures of the orphan or adopted orphan NRs were crystallized, a less prominent movement of H12 was apparent in these structures and challenged the status of the previous model as a universal NR activation model (Nolte et al., 1998).

The finding made in this study (I) that PPARs associate, in the absence of a ligand, with CoAs, could have two possible explanations: either an endogenous ligand (such as a fatty acid) present in all cells leads to a constant active state, or the structural features of the receptor allow the stabilization of H12 already in the absence of a ligand. This same question applies to other receptors that have been described to have constitutive activity, such as CAR, HNF4α, LRH-1, ERRγ, or the ROR isotypes. Of these, the crystal structure of HNF4α showed a constitutively bound lipid in its ligand-binding pocket (Wisely et al., 2002). Additionally, ROR isotypes were co-crystallized with cholesterol sulfate and retinoic acid (Kallen et al., 2002; Stehlin-Gaon et al., 2003).

Some of these known constitutively active NRs can also be modulated by ligands, as was described for LRH-1 and CAR (Krylova et al., 2005; Maglich et al., 2003). The role of the ligand, however, seems to be different in these receptors, since those receptors that could be crystallized without a ligand do not significantly differ from the ligand-bound structures, with respect to H12 position. This situation applies also to PPARγ and PPARβ/δ. The critical distance of the charge clamp residues (one of which is located on H12) for efficient CoA interaction was described to be 19 Å for the ligand-bound VDR (Väisänen et al., 2002). This distance is maintained in the apo-structures of LRH-1 (19.63 Å), CAR (18.3 Å) and PPAR isotypes (19.9 Å). This suggests a model, where the ligand may play a role in the additional stabilization of the active conformation in these receptors, but is not absolutely required for it. Furthermore, it could be argued that constitutive versus ligand-dependent activity has not diverged completely for the large majority of NRs. A model about NR evolution suggests that ligand-binding is an acquired property in the family (or conversely a loss of constitutive activity), which has developed most in the endocrine receptors that have lost entirely the property to be activated in the absence of a ligand (Bertrand et al., 2004).
The explanation of the ligand-independent CoA association is shown here to be based on the structural features of PPARs. However, further activation by ligand-binding is not excluded by this constitutive activity. Firstly, the structural comparison with constitutively active receptors suggested that the apo-forms of the PPAR isotypes represent receptors in active conformations, with respect to positioning of H12. In agreement, supershift and transactivation assays support the conclusion that the active apo-conformation enables CoA interaction. To further rule out the possibility that this effect is due to a natural ligand, present in the cell extracts, the overexpression of a CoR was shown to reduce the basal activity. This effect could be overcome by the addition of a synthetic ligand, which suggests that a strong natural ligand is not responsible for the high activity level seen in untreated cells.

6.2 Structural determinants of H12 stabilization in PPARs

Four groups of residues were shown by mutagenesis to contribute to the stabilization of H12. They modulate the PPAR protein surface at the CoA interface. Part of them do so directly, such as the charged surface residues and part indirectly, such as the internal residues that contact and stabilize H12 (i.e. without themselves being located at the protein-protein interaction surface). The charge clamp residues K329 from H3 and E499 from H12 in the first group are highly conserved and their role has been established for a number of NRs. Their exact positioning creates selectivity for the otherwise hydrophobic surface. A reversed charge clamp, observed in the NR Nurr1, is not known to interact with any known CoAs (Wang et al., 2003b). Our results are in agreement with previous studies highlighting their importance: a dramatic loss of both ligand-dependent and -independent activity was seen upon mutation of either residue. The interaction between charged residues of the second group, K347 and D503, conserved in mouse and human CAR (Frank et al., 2004), forms the first direct docking point in H12 stabilization and contributes to CoA stabilization by an additional direct contact. In the absence of ligand, mutations severely affected basal activity, reflecting losses in H12 stabilization.

The third group, consisting of the charged residues E352, D424, R425 and Y505 of H12, forms a link between the CoA surface and the other important protein-protein interaction surface, the heterodimerization interface. The contact to H12 (via Y505) is unique to PPARs. Finally, residues of the fourth group H351, H477, Y355 form a hydrogen-bonding network that encompasses Y501 from H12. Y501 was previously described as the ligand-contact point and shown to be crucial for ligand-dependent activation (Gampe et al., 2000). Here we show that the mutation of this residue does not alter ligand-independent activity as dramatically, an observation further against the hypothesis that an endogenous ligand would explain the observed constitutive activity of PPARs. Addition of a ligand further extends the hydrogen-bond network, yet direct H12 stabilization exists already in absence of a ligand. These data show that the inherent structural features of PPARs enable ligand-modulated activity as well as contribute to ligand-independent activity.
The structural permissivity of PPARs towards ligand-independent CoA association, combined with their rather ubiquitous expression pattern, suggests that the activity of these receptors may be more dependent on co-regulator protein concentrations or further structural modifications than previously thought. The ubiquitously expressed PPARβ/δ is probably the most ligand-sensitive member, having the least affinity for CoAs in a comparative panel and exhibiting a stronger affinity for CoRs. PPARγ2 expression on the other extreme is mainly differentiation-induced in adipose cells, where a ligand may modulate the signaling initiated. However, it could be argued that later the presence of the receptor may be sufficient to maintain expression of genes characteristic of adipocyte phenotype. Subtle daily changes in energy metabolism in response to food intake may be mediated by the modulation of receptor activity by other means, such as covalent modifications (for example through insulin/glucagon signaling). The specific responses that enhance their signaling through generation of ligands, may reflect an additional ligand-mediated level of the modulation of the gene expression profiles. Definite answers are lacking in this respect that encourage further research focusing on the comparison of these PPAR activation mechanisms.

6.3 Identification of the direct VDR and PPAR target genes of the IGFBP gene family

IGFs comprise a complex regulatory system that consists of two growth factors (IGF-I and IGF-II), two receptors (IGF-I R and IGF-II R), IGFBPs, four low-affinity IGFBPs (IGFBP-7, CTGF, NOV, IGFBP-10), proteases that modify the binding proteins and, in addition, several interacting molecules (Stewart and Rotwein, 1996). IGF-I and IGF-II are both growth-promoting growth factors, secreted by a variety of cells that circulate at concentrations higher than most other peptide hormones (e.g. 1000-times higher than insulin) (Daughaday and Rotwein, 1989). The initial interpretation of the function of the binding proteins was that they would prolong the serum half-life of the IGFs and modulate the tissue availability of the growth factors, since IGFBPs bind IGFs with higher affinities than the IGF-Rs. Several in vitro experiments and data obtained from transgenic animal models indicate that IGFBPs are more than just simple binding proteins (Kelley et al., 1996; Schneider et al., 2000). These studies show that IGFBPs are secreted in a tissue- and developmental stage-specific manner and they can modulate IGF bioactivity either positively or negatively. In addition to their role in IGF signaling, the IGFBPs have also been reported to function in some situations IGF-independently. These effects are to date less well characterized. Both IGFBP-3 and IGFBP-5 contain a nuclear localization signal (Radulescu, 1994; Schedlich et al., 1998). In addition, both have been also shown to mediate apoptosis, independent of IGFs (Rajah et al., 1997; Thomas et al., 1998).

Identification of binding sites is a prerequisite for the direct regulation of gene expression. In studies II and III, in silico RE detection methods based on in vitro binding strength data were used in the characterization of the regulation of IGFBP gene family members by VDR and PPARs.
6.3.1 Regulation of the IGFBP gene family by VDR

Previously, the IGFBP-3 gene had been reported as a direct VDR target gene with a DR3-type VDRE (Peng et al., 2004). A substantial body of literature has implicated the induction of the IGFBP-3 gene and its protein product by several other growth inhibitory proteins and agents including the tumor suppressor protein p53, retinoic acid and TGFβ (Buckbinder et al., 1995; Martin and Baxter, 1991; Shang et al., 1999). This suggests that this gene belongs to the set of genes that mediate the anti-proliferative effects of VDR. The direct regulation of this gene was confirmed here and additionally its direct genomic neighbor, IGFBP-1, was found to be a direct VDR target. The analysis of translation-independent and early-time scale regulation of IGFBP genes, demonstrated the direct regulation of yet a third family member, IGFBP-5. In contrast, IGFBP-2, IGFBP-4 and IGFBP-6 mRNA levels were unaffected. We observed these effects in bone and prostate cancer cells, but it is highly likely that these three genes are direct targets in other 1α,25(OH)₂D₃ responsive tissues. The 2- to 4-fold inductions observed are in the order of what was observed for most other 1α,25(OH)₂D₃ target genes.

IGFBP-5 resembles IGFBP-3 in its mechanisms of action. Both a secreted form and a cell surface interacting form that promotes IGF signaling due to the lowered affinity to IGFs have been described (Mohan et al., 1995; Twigg and Baxter, 1998). In addition, IGF-independent pro-apoptotic effects have been reported and the protein contains a nuclear localization signal as mentioned earlier. The main physiological roles of IGFBP-5 are in the development of the central nervous system, involution of the mammary gland (pro-apoptotic) and bone physiology (mitogenic) (Lee et al., 1995; Richman et al., 1999; Tonner et al., 1997). Since VDR plays a major role in bone homeostasis, the regulation of this family member fits well its regulatory domain.

IGFBP-1, on the other hand, has quite a distinct physiological role. It has been linked with several nutrition-related diseases. Insulin is known to regulate the synthesis of IGFBP-1 and post-prandial increases in serum insulin levels result in a corresponding four- to five-fold decrease in IGFBP-1 concentration (Busby et al., 1988). A study with non-insulin-dependent diabetes mellitus (NIDDM) patients, who do not have reduced insulin secretion, suggested that reduced IGFBP-1 levels correlated with risk factors for type II diabetes including impaired glucose tolerance, raised systolic and diastolic blood pressure, raised body mass index and raised serum triglyceride level (Heald et al., 2001). Decreased IGFBP-1 levels have also been added to the list of cardiovascular risk factors and have also been shown to predict risk of cardiovascular disease in type-2 diabetes. As discussed below, this gene is a shared target with PPARs. Although the role of VDR in the metabolic disorder has not been as widely characterized as that of PPARs, some shared target genes exist. For example, overlapping roles in the regulation of Insig genes, as described earlier, have been characterized.

The in silico approach applied here, resulted in the successful identification of multiple functional VDREs in the regulatory regions of the IGFBP-1, IGFBP-3 and IGFBP-5 genes, including ER-type elements (ER7), for which formerly no natural occurrences had been shown. Additionally, the ER9-type RE13 is one of the first natural elements of this type characterized. From the, in total, 174 kB...
genomic sequence, 15 candidate VDREs were identified. ChIP assays indicated that 10 of these regions are bound by the VDR in PC-3 cells, which represents a 67% success rate of the in silico binding site prediction. Moreover, the non-ligand responsive members IGFBP-4 and IGFBP-6, did not contain any VDRE candidates within 24 and 15 kB analyzed, respectively. The correlation of the in silico result with the mRNA inductions was not as straightforward in the case of IGFBP-2, where two intronic VDREs were identified. The experimental ChIP data was also not conclusive, both regions associated with VDR ex vivo. However, no VDR binding was detected on IGFBP-2 TSS by ChIP, which indicates that these VDREs may more probably be involved in the regulation of the neighboring IGFBP-5 gene. Interestingly, the intronic ER9-type VDRE of IGFBP-5 was found to contain a SNP that increases the spacing between the hexamers by 2 nucleotides. This significantly decreases its ability to function as a VDRE, as supported by the in vitro data. Unfortunately, the frequency of this SNP is very low (3%) and no cellular model was found to evaluate its significance in vivo.

The results concerning VDRE locations are in agreement with the whole genome analysis of conserved NR half sites that were found equally distributed around the TSS. Further, the in silico screening here was not restricted to cover only a few kB of proximal regulatory sequence (as in a lot of other studies), in effect several functional distal VDREs were identified. Based on the present understanding of enhancers, where DNA looping can occur free within chromatin units flanked by insulators, these distances are no limitations. However, such large distances between RE and TSS cannot be assessed by old-fashioned approaches, such as reporter gene assays with truncated constructs. Fortunately, more up-to-date ChIP assays allow the analysis of these distal regions.

The observation that the IGFBP-1/IGFBP-3 gene pair locus contains six functional VDR associated regions (2, 3, 4, 5, 6/7 and 8) and the IGFBP-5 gene area four such regions (9, 10, 13 and 14) supports the model of multiple REs per primary target genes, established earlier for the genes CYP24 and cyclin C (Sinkkonen et al., 2005; Väisanen et al., 2004). Furthermore, in a recent report, the autoregulation of VDR was mediated by two intronic VDREs (Zella et al., 2006), supporting both the lack of proximal bias for functional VDREs and the presence of several functional VDREs per target gene.

6.3.2 Regulation of the IGFBP gene family by PPARs

In this study, the main focus was on the IGFBP-1 gene, since this gene, from a physiological point of view, fits best the niche, where regulation by PPARs has been established in. We could demonstrate up-regulation of IGFBP-1 mRNA in human hepatocellular cells and in mouse liver during an early induction study (2 to 6 h) by all subtype selective PPAR ligands. This data shows that even though the three subtypes had differing expression levels, all of them were equally potent in inducing IGFBP-1 expression. Furthermore, even though the induction was not more than 2- to 3-fold and further considering the high basal expressed level of this gene, the effect is comparable to or even more substantial than that of majority of PPAR target genes, such as CPT1 used as reference here. It is also possible that PPARs may participate in the maintenance of the basal expression of this gene, as
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes

demonstrated by their presence at the promoter, already in absence of a ligand. Taken together with the previous observation that PPAR isotypes possess constitutive activity, it is possible that these NRs are acting in a positive way on a number of target genes that are yet to be discovered. The development of antagonists (direct or inverse) would be of great benefit for the discovery of these genes, by suppressing the constitutive activity of the PPARs.

In support of the likely direct regulation, putative PPREs were identified from a 20 kB region centered around the human IGFBP-1 gene TSS and two were shown to contribute to the regulation in transient transfection assays. Of these, the -1.2 kB element was most potent and additionally demonstrated a full association panel with the set of regulatory proteins tested in ChIP (PPARα, RXRα, p-PolII, TRAP220, NCoR1, RAC3 and PGC-1α). The fact that RE1 showed a more restricted association panel in ChIP may reflect weaker association to this element that may not result in sufficient signal levels for detection. Alternatively, the role of this region differs from that of RE2 and a different set of nuclear proteins are recruited.

The in silico search method proposed here differs from the alignment-based methods commonly in use, such as position-specific WMs, that were also recently applied in a genome-wide PPRE screen (Lemay and Hwang, 2006). These methods create mathematical scores over the whole length of the binding sequence, which makes them prone to overcompensate for nucleotide combinations that are completely unfavorable to TF binding at one position with scores from other positions. This problem arises especially in the scoring of larger motifs, such as the NR REs composed of two half sites. Therefore, we characterized the binding preferences of the PPARs to individual nucleotide variations of the consensus and used this data together with a set of experimentally verified binding results, to develop a set of rules how the half site variations can be combined without losing the affinity for the PPAR-RXR heterodimers completely.

This approach resulted in five putative REs in the human IGFBP-1 gene regulatory region, of which all were shown to bind in vitro at least one subtype. The binding strength of these elements was weak compared to the known CPT1 PPRE, including the functional REs 1 and 2. Regarding the in silico method, the successful identification of a set of PPREs that all associate with PPARs in vitro and identification of 2 of 5 functional REs demonstrates good selectivity and sensitivity (sequences down to the level of 1% binding could be identified). On the other hand, the low affinity of RE2 raises the question, whether PPAR binding to this site may require or be supported by the presence of other TFs. A sequence alignment of the human and mouse IGFBP-1 genes indicated that the sequence around this element showed a much higher conservation than regions of the four other REs. The occupancy of the IGFBP-1 gene TSS by PPAR isotypes was shown in HepG2s in presence or absence of activating ligands, yet was absent in HEK293 cells that do not express this gene. This could suggest that the region around RE1 functions as a selective enhancer module that requires presence of more factors than PPARs to get activated and induce basal and PPAR activated expression of the IGFBP-1 gene.
Discussion

The approach for detecting PPAR target genes that successfully identified IGFBP-1 as a direct target gene i) demonstration of an early induction profile, ii) presence of putative PPREs and iii) detection of PPAR binding on the TSS region by ChIP was applied to the other IGFBP family members. Those members, IGFBP-2, IGFBP-5 and IGFBP-6 that contain several putative PPREs were found to be responsive to PPAR ligands in HepG2 and HEK293 cells in an early time-course (2 to 6 h) and PPARs were shown to associate with their TSS regions after 2 h. In contrast, only the IGFBP-6 gene PPRE was identified with the alignment-based genome wide PPRE screen. Further analysis of candidate PPREs will be needed to fully classify these genes as direct targets, however, the data shown here supports this conclusion.

In addition, further investigations would be required to fully understand the physiological outcome of this regulation. IGFBP-1 may mainly have a role in connection to insulin signaling, rather than in control of cell growth, though the mechanisms of its action in metabolic control are largely unknown. The classical control of IGF signaling was demonstrated for this member in a mouse model, where it was shown to reduce hepatic neoplasia. The effect on cell proliferation of the other PPAR responsive family members may reflect more the tendency to promote the mitogenic IGF signal. The overexpression of IGFBP-2 was shown to be associated with increases in malignant growth and metastatic ability in cancer studies (Lee et al., 2005a). In addition, IGFBP-2 has been shown to be important as a bone anabolic signal bound to IGF-II. The two molecules form a binary complex that can traverse the capillary membrane (Arai et al., 1996) and thus, are available to extra-vascular tissues, preferentially skeletal tissue. Patients with the rare syndrome of Hepatitis C Associated Osteosclerosis (HCAO), who exhibit remarkable increases in bone mass as adults have been shown to have an elevation in IGF-II precursors and IGFBP-2 in their circulation (Khosla et al., 1998). IGFBP-6, found down-regulated here, is unique in that it predominantly binds IGF-II and inhibits its function by blocking its binding to the receptors (Roghani et al., 1989). In accordance, it has been shown to function as an autocrine growth inhibitory factor in human keratinocytes (Kato et al., 1995). In effect, the regulation pattern observed suggests a potential enhancement in IGF-II signaling. In the context of cancer, this effect is not wanted. Such conclusions based on conditions, where normal control mechanisms are compromised may not reveal what role this regulation has under normal physiological conditions. The dual nature of IGFBP-5 was discussed earlier in connection to repair of bone fractures where its role is to promote the mitogenic IGF signal and in the apoptotic process of mammary gland involution. The promotion of IGF signaling and the down-regulation of inhibitory signals may serve a beneficial functional role under certain physiological conditions, such as in injured tissues, which could link these genes to wound healing process that was described in connection to PPARβ/δ.

In conclusion, VDR and PPARs were shown to have both overlapping and diverging roles in the regulation of the IGFBP gene family. An interesting link between these two regulatory systems was recently shown by Dunlop and colleagues (Dunlop et al., 2005). This study reported an up-regulation of the human PPARβ/δ gene transcription by 1α,25(OH)2D3 and its receptor, VDR. Principles governing target gene regulation of these receptors showed important similarities with respect to

correlation between *in silico* predictions of putative binding sites and regulation that may serve useful for the identification of novel target genes, as demonstrated here among *IGFBP* gene family.

### 6.4 Meta-analysis of the binding site composition of human PPAR target genes

The pleiotropic actions of PPARs suggest that they regulate a large variety of target genes. This is supported by their wide expression pattern (especially that of PPARβ/δ). The acquisition of a full list of target genes as a collection of experiments from all these different tissues is challenging and further complicated by the possible differences in experimental model organisms and humans. In this study, we attempted to reach an understanding of common sequence features, the binding site composition, of PPAR target genes to offer an alternative approach to screen for potential targets and simultaneously to identify their regulatory elements. At the start of the study a relatively large collection of PPAR target genes with characterized REs existed, however with lack of binding strength data. The locations of these PPReS also suggested a positional bias towards the proximal promoter and many studies relied on conservation of PPReS.

To assess these parameters in an unbiased manner we started with a method to predict PPAR binding strength. We selected eight PPAR target genes, including both up-regulated and down-regulated targets, as well as varying PPRe locations. These genes were screened for PPRe of different binding strengths and for the gene *UCP3* a very close match of predicted and experimental binding strength was shown. This gene had no previously characterized PPRe and in this study we show three functional regions, including a strong distal PPRe. Overall, from these eight genes we identified several functional regions among the selection of PPReS made initially (we started with a selection of 10 proximal, 10 distal upstream and 10 distal downstream PPReS), mostly including the published PPReS but also suggesting novel regions. However, we could not confirm the functionality of the DR2- type *RVRα* PPRe. We also show that the human *ACOX* PPRe that was initially suggested as the regulatory region of this gene, is not functional, but instead this gene has two intronic PPReS. In summary, our data indicates no positional bias or a preferential use of conserved PPReS. Instead, we found a strong PPRe in each of these genes.

The analysis of additional 30 human PPAR targets allowed the extension of these conclusions based on the observed PPRe composition of the human gene and its mouse ortholog. A large majority of these genes in both human and mouse contained strong PPReS. PPReS could be found upstream and downstream of the genes, with only minor bias towards the proximal promoter. Preservation of PPReS between mouse and human was evident for some genes, yet even these genes often contained additional novel species-specific PPReS. In comparison to the published PPReS of this set of 38 targets, some genes seem to contain more PPReS, or stronger PPReS, than what was characterized earlier. These alternative sites should be validated further, however the high success of validation of strong PPReS in the eight genes characterized in detail suggests that these are most likely contributing to the responsiveness of the gene. Furthermore, those genes that contain a high
enrichment of PPREs that seems to have arisen in only the other species would offer an interesting test set to find out how many of these novel sites have acquired functionality.

There is also one interesting subset of genes that have a poor PPRE content in one species with very weak conserved sites that seem to form a stronger cluster in the other species (a retain-loss pattern). The IGFBP1 gene is a good example of such an exception where prevalence of medium or strong PPREs is only seen in the other species. Among the regions conserved between human and mouse, our previous study (III) showed the most proximal region to be functional. The cluster seen in mouse further upstream and its conserved human PPREs were not analyzed further, however the detailed conservation analysis carried out here suggests that also this region may be of interest for future studies. It still remains to be studied how such weak PPREs as seen in the human IGFBP1 can compete for the binding of the receptor and achieve transactivation. One likely explanation is that in their genomic context these PPREs function in a module of TFs inside which the cooperative binding helps attract and stabilize the receptor. The relatively narrow tissue distribution of this gene could be seen as an indication that this gene may indeed integrate the signaling of several TFs for its activation that may not be present in all tissues.

The analysis of two genes studied in detail was extended from the two-species comparison to a multi-species comparison. The genes ACOX1 and ANGPTL4 were selected since they represent two extremes, one with a very divergent PPRE composition versus another with a well-preserved one, respectively. The two most distant species to human, zebrafish and chicken had little in common (in terms of PPRE pattern) with either the human or the mouse gene, except that especially in zebrafish an abundance of strong/medium PPREs was seen in both genes. Surprisingly, the closest species to human analyzed, the chimpanzee, had also diverged from the human PPRE set. One functional PPRE in both ACOX1 and ANGPTL4 had been lost or weakened. While in mammals the ANGPTL4 gene was rather well-preserved, the best match to human seen in the dog gene, also this gene demonstrated that turnover of PPREs is occurring and novel sites seem to appear rather often, as seen in the rat gene as an example. Taken together, evolutionary constraints to maintain responsiveness do not translate to the preservation of an identical PPRE pattern. Instead, the appearance and fixation of novel site adds flexibility. Based on the set of human PPAR targets, this appears to manifest as an enrichment of strong or medium binding sites. The tracking of this enrichment and turnover from multiple species adds power to the in silico approach and enables detection of divergent PPRE patterns. Currently the mouse genome that has hardly any gaps serves as the best references, the extension to other species awaits the filling of the remaining gaps in their sequences.

As an example, the most gene dense human chromosome, chr 19, was screened for medium and strong PPREs together with its syntenic regions from the mouse genome. We presented the list of genes that resemble the known PPAR target set in both species and a more limited list of genes that show high enrichment and appearance of strong PPREs in human (to detect also human-specific targets). In total 11.2% of genes with orthologs were predicted as targets. The requirement for high enrichment and strong PPREs was filled by 8.2% of all known genes on chr 19. Together, these numbers suggest that approximately 10% of genes in the whole genome have features typical of
PPAR targets, which translates to a few thousand targets in total. This can be considered a relevant number based on microarray studies. Among the predicted targets, a number of genes with relevant physiology were identified, such as genes involved in lipid metabolism, regulation of cell cycle, mitochondrial transport and metabolism and inflammation. We selected six genes from these lists for further validation and could show that all responded to PPAR ligands. These included CYP4F8 involved in ω-oxidation, the NR COUP-TFγ, PAK4 involved in p21-signaling, the inner mitochondrial membrane translocase TIMM13, the fatty acid transporter SLC27A5 and LASS1. The LASS1 gene that participates in ceramide synthesis was studied in detail, which demonstrated that the predicted strong PPREs indeed are functional in reporter gene assay and ChIP.

In conclusion, this study showed that evaluation of binding strength is useful to discover target genes. Combined with the analysis of binding site composition from multiple species, this approach can reveal binding site turnover-patterns. In the chr 19 example we focused on predicted target genes that show typical features in both species. However, the analysis of multiple species can also suggest the regulation of genes where the binding site pattern has weakened in one species. This can be explored further as more complete genomes become available. These insights translate to a more efficient characterization of PPAR target genes and can be explored in the context of other TFs as well.
7. Summary and conclusions

In conclusion, the four scientific papers, on which this thesis is based on, represent the application of in silico methods in the analysis of the gene regulation by NRs, on protein and DNA level. On the protein level, a detailed investigation was made about functional properties directing cofactor-LBD interactions of PPARs. Whereas at the DNA level, the predictive power of in silico binding site search was evaluated for the identification of target genes of both VDR and PPARs.

Structural explanation for ligand independent CoA interaction in the human PPARs leading to the recognition of their constitutive activity (I)

We demonstrated that structural features of human PPARs allow ligand-independent CoA association comparable with the NR CAR. Using PPARγ as an example, we found that four different amino acid groups contribute to the ligand-independent stabilization of H12 of the PPAR LBD. These are: (I) K329-E499, mediating a charge clamp-type stabilization of H12 via a CoA bridge; (II) K347-D503, interacting with each other as well as contacting the CoA; (III) E352, D424, R425, R471 and Y505, directly stabilizing the H12 via salt bridges and hydrogen bonds; and (IV) H351, Y355, H477 and Y501, forming a hydrogen bond network. These amino acids are conserved within the PPAR subfamily, suggesting that the same mechanism applies for all three PPARs. Taking together, the ligand-independent contacts to H12 allow its positioning in active conformation allowing the interaction with CoA proteins. Further stabilization with a ligand is modest compared to that of VDR and similar to that of CAR. Therefore, PPARs more likely should be considered as active NRs in the absence of agonist and their functional profile classifies them close to the group of NRs with constitutive activity, such as CAR, RORs and LRH-1. The recognition of the constitutive activity of PPARs provides an additional view on PPAR signaling.

Regulation of the IGFBP gene family by VDR (II)

In this study we aimed to characterize the direct VDR targets of the IGFBP gene family. We found that IGFBP-1, its genomic neighbor IGFBP-3 and IGFBP-5 were regulated in a translation-independent manner in prostate and bone carcinoma cells and identified several functional VDREs in their vicinity. The VDRE in silico search based on previous experimental evaluation of published VDREs proved to be 67% efficient in predicting functional VDREs. This is a significant improvement compared to previous methods. Furthermore, the analysis of putative VDREs revealed a correlation between number of found REs and mRNA inductions: IGFBP-4 and IGFBP-6 contained no putative VDREs. IGFBP-2 had two distal VDREs in its intron but no inductions were observed in mRNA levels and no VDR associated at its TSS region. The results demonstrate that functional VDREs could be located distally to the TSS, both up- and downstream. Further, the importance of
Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes

natural ER-type VDREs was supported. The induction patterns of the IGFBP gene family provide insights into the anti-proliferative effects of 1\(\alpha\),25(OH)\(_2\)D

Regulation of the IGFBP gene family by PPARs (III)

This study shows that IGFBP-1 is a direct PPAR target in human hepatocarcinoma cells and mouse liver and contains two functional PPREs. Variations from the consensus PPRE were generated and evaluated by gelshift assays to build a PPRE-search tool. This in silico approach was further applied to the other family members and based on putative PPRE content IGFBP-2, IGFBP-5 and IGFBP-6 were predicted to be responsive. Their regulation was confirmed in human hepatocarcinoma and embryonic kidney cells: both mRNA inductions and association of PPARs on their TSS regions were observed. The functional IGFBP-1 binding sites are weak-affinity sites for PPARs, which most likely associate with other TFs in genomic context for further stabilization. This prompts the inclusion of surrounding sequence evaluation for in silico binding site prediction and emphasizes the importance of detailed characterization of binding preferences to cover the full set of possible target sequences. The existence of shared targets with VDR indicates that additional overlap in the regulomes may exist and emphasizes the role of this gene family in NR signaling.

Meta-analysis of known PPAR target genes

The initial findings concerning correlation of number of binding sites and mRNA inductions, based on the detailed analysis of the IGFBP gene family, were further evaluated, together with other parameters governing binding site functionality. An extended in silico analysis of PPAR regulated genes in human was performed, together with an experimental analysis of selected regions from eight of these genes. These findings resulted in a formulation of target gene prediction method emphasizing two of these variables: the binding strength of PPREs and enrichment of target gene regulatory areas in medium or strong binding PPREs. An initial test successfully identified both known and novel PPAR targets from human chr 19 and its syntenic regions from mouse genome. Furthermore, the results support the model of stabilizing selection acting on PPAR binding sites, where appearance of de novo PPREs can lead to species-specific binding site composition without the loss of regulation of the respective expression levels. These insights translate to a more efficient characterization of PPAR target genes and can be explored in the context of other TFs as well.
8. Future aspects

During the last decades bioinformatics has developed into an important research field for studying gene regulation. In parallel, molecular biology has advanced in methodology to collect data in a high-throughput fashion as is evident from the large collection of fully sequenced genomes, diverse datasets of genome-wide level expression analysis and more recent efforts to map regulatory proteins with ChIP from the unique sequence of the human genome. The impact of both is also seen in the NR research field.

To date many NR crystal structures have been determined. These have provided valuable information concerning structural differences that translate into functional diversity within the family. A large collection of LBDs exists, with only some orphan NRs lacking. On the other hand, less focus has been on the NR DBDs. However, a more diverse collection of DBD structures would allow the modeling of NR binding to consensus sequence variants to improve binding strength predictions. Moreover, a crystal structure of a whole receptor is still lacking, which will enable the full analysis of NR structural dynamics.

From the in silico analysis of TF binding sites it has become evident that appreciating as much in detail as possible the parameters that govern TF binding site functionality has the potential of improving search data in silico and will also help in experimental design to study the effects in vivo. This requires a good understanding of the binding site sequence variations tolerated by each TF, proper alignment methods to detect conservation, means to predict clusters of TF sites where less stringent affinity parameters may apply, appreciation of tissue selectivity of regulation and possibly means to predict whether predicted targets fit the physiology of the TF. Such analysis will greatly benefit from further collection of large datasets of changes in expression levels with a good coverage of responses and whole genome binding data for different TFs.

Equally challenging will be the interpretation of the results in terms of co-regulation patterns and higher levels of interactions inherent to complex systems. The dimensionality of this data poses a challenge for interpretation. Long lists of regulated genes, listed from hundreds of tissues, even with explanation of their biological role, are difficult, or even impossible to digest. This data is readily processed by a computer, therefore the development of bioinformatics tools for the visualization and interpretation of these datasets promises to simplify the task. Taken together, the combined efforts of systems biology modeling, bioinformatics predictions and experimental validations will be seen in future regulome research. In effect, each researcher in the bioscience field will also need to integrate the use of computational resources together with lab bench work.

In the NR field regulome level research promises to reveal important aspects of endocrine and nutrient signaling that in future hopefully can help prevent and cure such diseases that develop later
Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes

in life, with continued signaling between the environment and the genome playing a prominent role in disease progression. The knowledge of regulatory DNA used by each receptor will enable NR signaling to be assessed on an individual level: the detection of variation in these DNA stretches in different populations and also within populations, promises to bring a personalized aspect to biology and medicine.
References


Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes


References


Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes


protein-1 (IGFBP-1) with glucose tolerance and cardiovascular risk in two populations. Diabetologia 44, 333-339.


References


References


References


References


Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes


Merja Matilainen: Identification and characterization of target genes of the nuclear receptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes


Appendix: Original publications
Kuopio University Publications C. Natural and Environmental Sciences


