KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

MOTIF DETECTION IN VERTEBRATES BASED ON COMPARATIVE GENOMICS

Jury:
Prof. dr. ir. H. Hens, voorzitter
Prof. dr. ir. B. De Moor, promotor
Prof. dr. ir. K. Marchal, co-promotor
Prof. dr. ir. J. Suykens
Prof. dr. Y. Van de Peer (U.Gent)
Dr. ir. L. Verlinden
Prof. dr. J. Winderickx
Prof. dr. ir. J. Vanderleyden
Prof. dr. A. Verstuyf

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Ruth VAN HELLEMONT

U.D.C. 681.3*J3 Maart 2007

D/2007/7515/20

ISBN 978-90-5682-786-1

Voorwoord

Nu ik de laatste hand leg aan wat me de voorbije maanden heeft bezig gehouden, wordt het eens dringend tijd om terug te blikken en te bekijken wie me allemaal heeft geholpen dit werk rond te krijgen. Hierbij denk ik aan mensen die me advies gaven in verband met onderzoeksgerelateerde zaken of typische doctoraatsbeslommeringen, mensen die de hele periode aangenaam hebben gemaakt of mensen die beide hebben gedaan: allemaal hebben jullie vandaag mogelijk gemaakt!

Eerst en vooral wil ik mijn promoter, Bart De Moor, bedanken omdat hij me de kans heeft gegeven mijn onderzoek uit te voeren binnen de bio-informatica groep van ESAT-SCD.

Kathleen, ik weet zelfs niet hoe ik moet beginnen met jou te bedanken… Eerst en vooral heb je me altijd weten te motiveren ook en vooral als ik het soms effe niet meer zag zitten (bijvoorbeeld na de derde review van het Genome Biology-artikel). Ik zal nooit vergeten hoe je, de maanden dat ik thuis was, regelmatig eens op de chat kwam polsen hoe het met me ging en me zo toch het gevoel gaf dat ik ergens nog een normaal leven had, met collega's die met me mee leefden. Ook in de moeilijke maanden die volgden op mijn terugkeer naar ESAT, heb je regelmatig op me ingepraat en me het vertrouwen gegeven om er weer in te vliegen. Ik denk dat net die interesse en dat vertouwen in je studenten, maken dat je een supergoeie 'baas' bent. Dat gecombineerd met je grappige uitspattingen en eigenaardige lollige trekjes, maakt dat iedereen een beetje in uw buurt blijft hangen … Hoewel ik de uitzondering ben die de regel bevestigt, hoop ik in de toekomst nog contact te houden, al is het maar om me op de hoogte te houden van bioi-roddels en elkaar bij te schaven op gebied van de laatste modetrends.

Yves, u wil ik hartelijk bedanken voor de toffe samenwerking en omdat je me de kans hebt te gegeven om te werken op subfunctionalizatie. Als ik op het einde van mijn doctoraatsperiode van iets spijt heb, is het dat ik niet meer tijd heb gehad om subfunctionalizatie genoomwijd te bestuderen. Ik ben echt benieuwd wat hieruit zal komen (Tine en Valerie, bij deze wat extra druk op jullie schouders ;-) en de vraag of jullie me op de hoogte willen houden). Wat ik vooral heel erg geapprecieerd heb, is dat je me steeds hebt behandeld als een van je eigen studenten. Ik kreeg steeds snelle, grondige en zeer enthousiaste feedback.

Guy, Lieve en Mieke, ook jullie wil ik bedanken voor de leuke samenwerking. Ik vond het geweldig van die abstracte motiefjes eens gevalideerd te zien. Dat maakt het allemaal iets echter! Verder ook een dikke merci om de vitamine D_3-hoofdstukken na te lezen. Lieve en Mieke, jullie wil ik in het bijzonder bedanken omdat jullie bovendien in mijn jury willen zetelen.

Johan, Joris en Jos, ook jullie wil ik hartelijk bedanken omdat jullie de tijd hebben genomen om mijn doctoraatstekst door te nemen, te beoordelen en hier vandaag als juryleden aanwezig willen zijn.

Ik wil ook het IWT bedanken, dat mij vier jaar lang financieel gesteund heeft, en zonder hetwelk dit onderzoek nooit was mogelijk geweest.

Mevr. Van Broeck, het is dankzij uw boeiende lessen over erfelijkheid dat ik besloot bio-ingenieur in de cel- en genbiotechnologie te gaan studeren. Het boekje dat hier nu ligt is daar een rechtsreeks gevolg van. Bedankt hiervoor.

Natuurlijk wil ik de collega's van de bioinformatica-groep hier ook vermelden: de geanimeerde lunches, de artikel-tractaties bij mama Wong, de talrijke recepties en geslaagde scd-weekends maakte het steeds fijn werken op ESAT! In het bijzonder wil ik de mensen bedanken die het dagdagelijks met me uithielden aan ons eiland: Pieter en Kristof, naast onze geslaagde '*expérience Tour Eifel*' vond ik jullie bijzonder leuke eiland-buddies. Kristof, merci voor de mini courses Matlab (die eigenlijk niet aan mij besteed waren) en de tokitokiboombooms. Je ging verrassend gemakkelijk om met de opeenvolgende nederlagen ;-). Pieter, ik denk dat jij zonder twijfel de collega bent die de meeste thx-credits heeft vergaard: hulp met debuggen van scripten, personal coach van het administratieve proces in de aanloop naar een doctoraat (miserie!), koffiepauze-maatje (vaak op uw 'kosten' ;-). Kortom merci voor alles: ik geef u een boekske en een koffiekaartje en we staan weer quitte :-).

Bert, mijn enige "felicitaties van de jury"-vriend, voor een burgie val je best mee :-). Rafke,, a.k.a. the one and only 'baby', desondanks al de warpholes in je brein vind ik u ne toffe gast. "Mannekes", jullie gaven mijn ESAT-carrière de grondige dosis humor.

Ook ontspanning is essentieel tijdens het schrijven van een doctoraat (en daarnaast) en daarvoor was ik goed omringd: Greet, Kathleen, Sofie en Wendy, bedankt voor de leuke tijden tijdens onze studies en daarna. Ik geniet altijd van onze bijeenkomsten en kom steevast opgeladen van energie terug. Ik vind het ook heel sympathiek dat jullie willen voordoen hoe een werkleven te combineren valt met een baby, zodat ik dat ook weet voor binnen enkele jaren ;-).

Femke, Ingrid, Jan, Toon en Veerle, onze x-maandelijkse etentjes in Leuven zijn altijd super grappig! Alleen de steeds beperktere groep resto's waar we nog met opgeheven hoofd binnen kunnen wandelen maakt het soms wat moeilijk… Telkens we elkaar zien, amuseer ik me rot!

Hilde, merci voor me regelmatig uit mijn kot te halen voor een gezellige en ontspannende avond in het Brusselse. Roos, ik vind het tof dat je terug in België bent, ik hoop dat we nog regelmatig eens kunnen afspreken.

Cathy en Jerry, ondanks het feit dat we elkaar maar sporadisch zien (misschien beter!) zijn het keer op keer geslaagde avonden. De kers op de taart wordt onze trip naar de Champagnestreek! 's Avonds met de taxi betekent "no BOBs no rules!"

Tony, Justine, Steve, Kimley, Julle, Silvia, Kevin, Corinne, Tim, Marijke en Stephane, 't prinstepolschte is dat we ons altijd goed amuseren! Steveke, leeftijdsgenoot, merci om me altijd enkele weken voor te bereiden op de aftakeling die me te wachten staat. Zo sta ik niet alleen met al mijn wijsheid in een groep van groentjes ;-). Tony en Justine, nen dikke merci voor de talloze decadente, hilarische en bijna altijd uit de hand gelopen etentjes. Ze waren noodzakelijk om te ontspannen maar 's anderdaags droegen ze niet bij tot de productiviteit. Ook onze vakantie in Italië was supergeslaagd en zal ik niet gauw vergeten. Ik wil jullie eveneens bedanken dat ik 'schaduw-meter' mag zijn van Lili, de allerliefste baby in de Brusselse contreien.

Peeters Management Consulting & Coaching, beter gekend als Jean-Marie en Chantal, heeft me al ettelijk keren consulting en coaching geven. Bedankt!

Mimi en Leo, ik vind het nog altijd *super* dat jullie speciaal een op-en-afje van skivakantie maken om er vandaag bij te zijn. Daarnaast wil ik jullie natuurlijk ook bedanken voor de talrijke lekkere etentjes en gezellige avonden. Jullie hebben een plusje voor op 'gewone' schoonouders. Elketje, bedankt voor de attente telefoontjes om te polsen hoe het met me ging.

Anna, je bent er gewoon *altijd* geweest, zolang ik me kan herinneren. Van pampers, over huiswerk, rebelse puberteit, examenstress tot nu. Je was er in elke nieuwe faze van mijn leven.

Omaatje, elke week kijk ik weer uit naar onze toffe telefoongesprekken, de laatste maanden misschien des te meer. Ik ben altijd benieuwd naar je belevenissen en grappige, eigenzinnig ingekleurde verhalen. Mamy, u wil ik bedanken omdat ik uw zonnetje mag zijn. Ik besef goed hoe trots jullie zijn op mij!

Ook de rest van mijn familie -tantes, nonkels, neven en nichten- bedankt om er altijd te zijn. Jef, u wil ik in het bijzonder bedanken voor het

nalezen van mijn Engelse tekst. Bericht aan tante Miet: “nog één doctoraatje en ik heb je ingehaald ;-)”. Guido en Andrée, jullie verdienen zeker een speciale vermelding voor alles wat jullie voor me gedaan hebben en betekenen. Telkens ik een probleem heb van welke aard dan ook, kan ik bij jullie terecht voor hulp, raad of gewoon een opluchtende babbel. En dat geeft een gerust gevoel. Bedankt!

Mama en papa hoewel jullie er spijtig genoeg niet meer bij kunnen zijn, weet ik dat jullie erg trots zouden geweest zijn. Papa, zou je ooit gedacht hebben dat je tegendraadse puberdochter van toen ooit een ‘serieuze’ doctor zou worden? Moepie, jij wist dit wel… en ook de chaotische manier waarop ik de laatste maanden heb gewerkt zul je waarschijnlijk voorspeld hebben. Ik heb dan ook je raad, rust en structuur gemist. Dank je voor het vertrouwen: “We kunnen inderdaad onze plan trekken”…

Lientertje, mijn kleine zus, de laatste maanden was je meer als een grote zus voor mij. Je stond altijd direct klaar voor mij bij de minste kreet van paniek (of aandacht ;-). Dank je voor het nalezen en verbeteren van mijn Nederlandse samenvatting alsook het nakijken van de proefdruk. Maar meer nog bedankt voor de leuke ontspanningsmomenten: lekkere bijklets-lunches, al dan niet gelinkt aan een zwempartij, portefeuille-legende shopnamiddagen en supergezellige familie-zondagavonden. En dat alles gecombineerd met een goede portie geklets en gezwans (al werd dat soms bruut getemperd op zondagavond :-). Je bent een super zus en samen met Jochen vorm je een meer-dan-volwaardig gezin voor mij.

Liefmans, terwijl ik hier de laatste hand leg aan dit dankwoord, ben jij om half-twee ‘s nachts vol enthousiasme (“****-ing Word!”) mijn inhoudstabel aan het maken terwijl je morgen een belangrijke presentatie moet geven bij je klant. En toch heb ik de laatste maanden soms in (de sporadische ;-) lastige momenten durven beweren dat je er niet steeds voor me was. Nu het einde in zicht is, kan ik heel eerlijk zeggen dat je de schrijfperiode stukken lichter hebt gemaakt door je supergrappige en lieve zelf te zijn en omdat ik weet dat ik altijd op je kan tellen (al moet ik soms wat geduld hebben ;-). Onder andere je Witse-dansjes, thuiskomsten met open-armpjes, Mowgli-imitaties (ja, je bent echt zo’n rariteiten-cabinet ;-) en versmachtende knuffels hebben ervoor gezorgd dat ik niet zot werd van dat eindeloze thuiszitten en schrijven. Dank je wel. Ik kijk er naar uit om samen aan een nieuw hoofdstuk te beginnen…

Ik weet nu al dat ik veel mensen ben vergeten te vermelden die evengoed een rol hebben gespeeld in de laatste jaren. Sorry … bij deze toch bedankt!

Ruth

Abstract

Vertebrate organisms consist of multiple cell types. Although each cell contains an identical amount of genetic information, each fulfils a different function within the organism. Moreover, each cell is able to adapt to changing environmental conditions, such as developmental stage, nutrient level, etc. Part of this flexibility is embedded in the changes in gene expression, controlled at the level of transcription. Therefore, in this thesis we attempted to unravel transcriptional regulation of genes by studying their regulatory motifs. Motifs are short DNA sequences located in the promoter region of a gene, which serve as recognition tag for the corresponding transcriptional regulators.

Many approaches have been developed to identify such regulatory motifs in an automatic way, such as motif screening and motif detection methods. With the availability of complete genomes comparative genomics can also be used to recover regulatory motifs, a strategy called phylogenetic footprinting. Although each of these methodologies proved to be successful to study certain biological problems, none was optimally suited to identify regulatory motifs in intergenic sequences of highly diverged vertebrates.

Therefore, in this PhD, we combined different existing methods in order to make use of their strengths and minimize their drawbacks.

We developed a new procedure for phylogenetic footprinting that combines alignment and probabilistic motif detection. This methodology proved to be successful in identifying evolutionary conserved regulatory motifs in highly diverged vertebrate organisms.

Furthermore, we developed a workflow to identify evolutionary conserved regulatory motifs that are divergently retained between fish paralogs, in concordance with the divergent expression profile of the fish duplicates, i.e., subfunctionalization.

Finally, we aimed at gaining more insight in the effects that are attributed to vitamin D_3. We started from groups of genes that show a similar expression profile after cells have been treated with vitamin D_3. We assume that this common expression profile is due to a common regulatory motif that controls the expression of the co-expressed genes. First, we investigated the role of two well-known regulatory motifs in the molecular mechanism behind the actions of vitamin D_3. We identified E2F as an important player in vitamin D_3-induced growth inhibition. Additionally we were able to

recover a number of previously unknown E2F targets, which might lead to a better resolving of the molecular mechanism behind the non-classical vitamin D_3 effects.

Furthermore, we developed a strategy to identify *de novo* reliable evolutionary conserved regulatory motifs that are possibly responsible for the observed co-expression of genes and we applied this methodology to a group of genes that are up-regulated after cells have been treated with vitamin D_3. This led to the identification of 31 motifs that are present in the intergenic region of multiple vitamin D_3-regulated genes. Further experimental work will be necessary to reveal their biological function.

All the methodologies developed and/or applied in this thesis are generic. They can be used to identify regulatory motifs in a wide spectrum of biological problems, ranging from a specific pathway (e.g., antiproliferative action of vitamin D_3) to a genome wide study.

Korte inhoud

Vertebrate organismen zijn opgebouwd uit verschillende celtypen. Hoewel elke cel dezelfde genetische informatie bevat, vervult elk celtype een verschillende functie. Daarenboven is elke cel in staat zich aan te passen aan veranderende omgevingsfactoren, zoals ontwikkelingsstadium, beschikbaarheid van nutriënten, enz. Deze flexibiliteit is grotendeels te wijten aan veranderingen in genexpressie die gecontroleerd worden door middel van transcriptionele regulatie. Het doel van deze thesis was om de transcriptionele regulatie te ontrafelen door het bestuderen van regulatorische motieven. Regulatorische motieven zijn korte DNA-sequenties, gelegen in de promoterregio van een gen, die dienen als herkennings- en bindingsplaats voor de overeenkomstige transcriptionele regulators.

Er zijn reeds verscheidene methoden ontwikkeld om op een geautomatiseerde manier regulatorische motieven te identificeren, zoals motiefscreening- en motiefdetectiemethoden. Door de beschikbaarheid van volledige genomen van diverse species, kan eveneens comparatieve genoomanalyse gebruikt worden voor de identificatie van regulatorische motieven; een strategie die *phylogenetic footprinting* wordt genoemd. Hoewel elk van deze methoden succesvol is gebleken voor het bestuderen van bepaalde biologische problemen, bleek geen enkele optimaal voor de identificatie van regulatorische motieven in de intergensiche sequenties van sterk gedivergeerde vertebraten.

Daarom werden in deze thesis de verschillende bestaande methoden gecombineerd, waardoor optimaal gebruik gemaakt werd van hun sterktes en hun beperkingen geminimaliseerd werden.

We hebben een nieuwe procedure ontwikkeld voor *phylogenetic footprinting* waarin alignering en probabilistische motiefdetectie samengevoegd worden. Deze methodologie bleek succesvol voor de identificatie van evolutionair geconserveerde regulatorische motieven in sterk gedivergeerde vertebrate organismen.

Bovendien, ontwikkelden we een methodologie voor het onderscheiden van evolutionair geconserveerde regulatorische motieven die differentieel behouden zijn tussen visparalogen en dit in overeenkomst met de expressieverschillen tussen beide paralogen (subfunctionalizatie).

We trachtten ook meer inzicht te verwerven in de werking van vitamine D_3. Hiervoor startten we van groepen van genen die een gelijkaardig expressiepatroon vertoonden nadat cellen behandeld waren met vitamine D_3. We veronderstellen dat zo een gemeenschappelijk expressieprofiel te wijten is aan een gemeenschappelijk regulatorisch motief dat de expressie van de co-gereguleerde genen controleert. Eerst bestudeerden we de rol van twee goedgekarakteriseerde regulatorische motieven. We stelden vast dat E2F een belangrijke rol speelt in de vitamine D_3-geïnduceerde groei-inhibitie. Bovendien identificeerden we een aantal voordien ongekende E2F-doelwitgenen.

We ontwikkelden eveneens een *de novo* strategie om regulatorische motieven te identificeren die mogelijk verantwoordelijk zijn voor het expressiepatroon van co-gereguleerde genen. Deze methodologie werd toegepast op een groep van vitamine D_3-gereguleerde genen met een gelijkaardig expressiepatroon. Dit leidde tot de identificatie van 31 motieven die aanwezig zijn in de promoterregio van deze co-gereguleerde genen. Om hun biologische rol op te helderen is nog additioneel experimenteel werk nodig.

Al de methodologieën die in deze thesis ontwikkeld en/of toegepast werden zijn generisch. Ze kunnen dus gebruikt worden voor de identificatie van regulatorische motieven in een breed spectrum van biologische problemen, van een specifieke reactieweg (bijvoorbeeld antiproliferatieve werking van vitamine D_3) tot genoomwijde studies.

Acronyms

A	adenine
B	C, G, T
bHLH	basic helix-loop-helix
bp	base pair(s)
C	cytosine
CDK	cyclin-dependent kinase
CDKI	CDK inhibitor
D	A, G, T
DNA	deoxy-ribonucleic acid
DP	differentiation-regulated transcription factor 1 protein, DRTF1 protein
DyP	dynamic programming
E2F	E2A binding factor
ECR	evolutionary conserved regions
EM	Expectation-Maximization
EPD	Eukaryotic Promoter Database
FSD	fish specific duplication
G	guanine
G1-phase	gap 1-phase
G2-phase	gap 2-phase
GO	Gene Ontology
GOLD	Genomes OnLine Database
GRN	gene regulatory network
H	A, C, T
INK4	inhibitor of CDK4
IUPAC	International Union of Pure and Applied Chemistry

K	G or T
kb	kilo base pair(s)
M	A or C
MCM	minichromosome maintenance
MHC	major histocompatibility complex
melk/Melk	Maternal embryonic leucine zipper kinase
M-phase	mitosis phase
mRNA	messenger RNA
mya	million years ago
N	A, C, G, T
NCBI	National Centre for Biotechnology Information
NISC	NIH Intramural Sequencing Center
nt	nucleotide(s)
NW	Needleman-Wunsch
PcG protein	Polycomb Group protein
PF	phylogenetic footprinting
PSFM	position specific frequency matrix
PWM	position weight matrix
R	A or G
Rb	retinoblastoma protein
RNA	ribonucleic acid
RNAPII	RNA polymerase II
RP	restriction point
S	G or C
SCPD	*Saccharomyces cerevisiae* Promoter Database
S-phase	synthesis phase
SW	Smith-Waterman
T	thymine
TAF	TBP-associated factor
TBA	Threaded Blockset Aligner

TBP	TATA-binding protein
TF	transcription factor
TFBS	transcription factor binding site
TGI	The Gene Index Project
Th1	T helper cell type 1
Th2	T helper cell type 2
tss	transcription start site
TUF	transcript of unknown function
UCR	ultra-conserved region
UTR	untranslated region
utss	upstream from tss
V	A, C, G
VDR	vitamin D receptor
VDRE	Vitamin D response element
vitD3	1,25-dihydroxyvitamin D_3
W	A or T
Y	C or T

Nederlandse samenvatting

Hoofdstuk 1: Inleiding

Situering van de thesis

Sinds de publicatie van het menselijk genoom in 2001, zijn reeds vele andere hogere eukaryote genomen in kaart gebracht. De GOLD databank bevatte in november 2007, 514 volledig gekende genomen, waarvan 47 eukaryote genomen. Deze omvatten 7 genomen van gewervelde dieren, met name de groene kogelvis, de tijgerkogelvis, mens, muis, rat, zebravis en de primitieve vertebraat zakpijp. Er zijn momenteel ook honderden genoomsequentieprojecten lopende, van o.a. amfibieën, bijvoorbeeld.kikker; vogels, bijvoorbeeld kip; vissen, bijvoorbeeld zalm en zoogdieren, bijvoorbeeld chimpansee, hond, kat, paard en varken. Recent werd daarenboven een gedeelte van het Neanderthaler DNA in kaart gebracht. Deze bron van genetische informatie kan onder meer gebruikt worden voor paleontologische doeleinden zoals de berekening van de divergentietijd tussen de mens en de Neanderthaler.

De toenemende hoeveelheid genetische informatie, die het gevolg is van deze recente sequeneringsinspanningen, wordt gestockeerd in genoomdatabanken waar ze beschikbaar worden gesteld van wetenschappers. De belangrijkste databanken, zoals bijvoorbeeld Ensembl, NCBI, NISC en TGI, verdubbelen ongeveer elke 18 maanden in grootte, als gevolg van de toenemende hoeveelheid gekende genoomsequenties.

De volledige genoomsequenties worden o.a. gebruikt voor comparatieve genomica: vergelijking van sequenties van verschillende organismen maakt het mogelijk sequentie-elementen te identificeren die aanwezig zijn in verschillende organismen en die dus geconserveerd bleven gedurende de evolutie. Deze geconserveerde elementen worden verondersteld functionele elementen te zijn, zoals proteïne-coderende genen of exons. De onderliggende redenering is de volgende: een toevallige mutatie in een functioneel sequentie-element is meestal funest voor het organisme en daarom is het onwaarschijnlijk dat zulke mutaties zich accumuleren in de tijd. Als gevolg blijven functionele elementen gespaard

van grote aantallen mutaties en blijven ze behouden in verschillende organismen.

Wanneer de genoomsequenties van mens en muis vergeleken worden, blijkt 5% behouden onder selectie druk. Gegeven dat proteïne-coderende elementen circa 1.5% uitmaken van het menselijk genoom, impliceert dat er sprake is van een ander type geconserveerde sequentie-elementen. Deze hypothese wordt ondersteund door de observatie dat intergenische (niet-proteïne-coderende) gebieden meer behouden zijn tussen organismen dan verwacht wordt van functieloos DNA; zo zijn de intergensiche gebieden van mens, muis en hond bijvoorbeeld voor 4% identiek. Dit toont aan dat zoogdiergenomen meer functionele sequentie-elementen bevatten, zoals onder meer regulatorische motieven.

Deze thesis behandelt de identificatie van regulatorische motieven op basis van hun conservatie profiel. In de volgende twee paragrafen wordt kort gedefinieerd wat regulatorische motieven zijn en wat hun rol is in de transcriptionele regulatie.

Transcriptionele regulatie in eukaryoten

Transcriptie is het biologisch proces waarbij DNA wordt overgeschreven naar RNA. De hoeveelheid mRNA (messenger of boodschapper RNA) afgeschreven van een bepaald gen dat op een bepaald ogenblik aanwezig is in de cel, noemt men de ‘expressie’ van dat gen op dat ogenblik. Voor de meeste proteïne-coderende genen varieert dit expressieniveau afhankelijk van de omstandigheden, zoals bijvoorbeeld ontwikkelingsstadium, celtype, aanwezigheid van nutriënten, enzovoort. Het expressieniveau van een individueel gen wordt het sterkst gecontroleerd ter hoogte van de transcriptie. Regulatie van transcriptie is een dynamisch proces waaraan veel verschillende factoren deelnemen: binding van RNA-polymerase en van de transcriptie-initiatiefactoren op de basale promoter initieert een beperkt niveau van transcriptie. Bovendien is deze vorm van transcriptie-initiatie algemeen aanwezig in alle celtypes en biedt deze geen regulatorische specificiteit. Om functioneel significante hoeveelheden mRNA te produceren is de sequentiespecifieke binding nodig van transcriptiefactoren op transcriptiefactor-bindingsplaatsen, gelegen buiten de basale promoter. Deze bindingsplaatsen worden ook regulatorische motieven genoemd.

Regulatorische elementen

Regulatorische motieven zijn korte DNA-sequenties van gemiddeld 5 tot 8 nucleotiden (nt), die zowel voorkomen in de promoter regio (d.w.z. het intergenisch gebied dat onmiddellijk stroomopwaarts ligt van de start van het gen) als op lange afstand van het doelwitgen, het gen waarvan ze de transcriptie beïnvloeden.

Regulatorische motieven worden sequentiespecifiek herkend en gebonden door een transcriptiefactor. Deze transcriptiefactor kan zowel een 'activator' zijn die de transcriptie van het doelwitgen bevordert, als een 'repressor', die de transcriptie vermindert. Transcriptiefactoren binden regulatorische motieven zodanig dat ze in de juiste positie staan t.o.v. andere transcriptiefactoren en t.o.v. het basale transcriptieapparaat. Transcriptiefactoren bepalen de snelheid waarmee het doelwitgen wordt afgeschreven: zij zorgen ervoor dat de expressie van een gen geactiveerd of gerepresseerd wordt in bepaalde omstandigheden.

Regulatorische motieven kunnen op verschillende manieren worden voorgesteld. De consensussequentie is de meest eenvoudige weergave en geeft voor elke positie van het regulatorische motief het meest voorkomende nucleotide weer (eventueel m.b.v. gedegenereerde symbolen). Een meer geavanceerde manier om een regulatorisch motief weer te geven is door middel van een matrix model: voor elke positie in het motief wordt de probabiliteit weergegeven waarmee een bepaald nucleotide op die plaats waargenomen wordt. Een derde representatiewijze is het motieflogo, gebaseerd op de matrixweergave. Hierbij wordt voor elke positie de frequentie van een specifiek nucleotide voorgesteld met zijn overeenkomstig symbool (A, C, G of T), waarbij de hoogte van het symbool evenredig is met de frequentie van het overeenkomstig nucleotide.

Een regulatorische module, is een groep van regulatorische motieven die samen gelokaliseerd zijn in het intergenisch gebied van een gen. Regulatorische modules kunnen variëren in lengte van enkele honderden basenparen (bp) tot meer dan 100 kilobasenparen (kb). Zij bevatten gemiddeld 6 tot 15 regulatorische motieven voor 4 tot 8 verschillende transcriptiefactoren.

Regulatorische motieven en modules vormen samen met de genen die ze reguleren de bouwstenen van regulatorische netwerken.

Computationele detectie van regulatorische motieven

Omdat de empirische validatie van regulatorische motieven arbeidsintensief is, werden verschillende computationele methoden ontwikkeld voor de identificatie van zulke bindingsplaatsen (regulatorische motieven). Over het algemeen kan men hierbij twee grote categorieën onderscheiden.

De eerste categorie van algoritmen vertrekt van gekende transcriptiefactor-bindingsplaatsen en zoekt in sequenties naar deze gekende motieven. Deze methodes worden *motiefscreeningmethoden* genoemd.

De tweede werkwijze zoekt naar regulatorische motieven zonder enige voorkennis over welke bindingsplaats precies gezocht wordt. Deze *de novo* methoden maken het mogelijk nieuwe, nog niet gekarakteriseerde motieven te ontdekken. Deze categorie van methoden omvat op zijn beurt twee strategieën. De eerste benadering is gebaseerd op de hypothese dat de expressie van genen die co-gereguleerd zijn vermoedelijk gecontroleerd wordt door hetzelfde transcriptioneel regulatiemechanisme, d.w.z. dat deze genen dezelfde regulatorische motieven dragen in hun intergenisch gebied. Dit zijn de *motiefdetectiemethoden*. De tweede aanpak is '*phylogenetic footprinting'*. Hierbij wordt gebruik gemaakt van comparatieve genomica: regulatorische motieven worden geïdentificeerd als sequentie-elementen die geconserveerd zijn tussen orthologe intergensiche sequenties. De onderliggende hypothese is dat evolutie de biologisch relevante sequenties bewaart, ook in intergenische gebieden.

Probleemstelling

Hoewel er reeds verschillende methoden bestaan om regulatorische motieven te identificeren, zijn deze geen van allen optimaal voor motiefidentificatie in sterk gedivergeerde vertebraten. In Tabel N-1 worden de belangrijkste nadelen van elke strategie (zie vorige paragraaf) opgesomd.

Omdat er slechts een fractie van de functionele transcriptiefactor-bindingsplaatsen gekend is, is de toepasbaarheid van *motiefscreeningmethoden* gelimiteerd. Deze methoden zijn vooral nuttig voor de studie van een of meerdere specifieke regulators, met name om alle mogelijke bindingsplaatsen voor deze regulators terug te vinden. *Motiefscreening* zal echter niet resulteren in de identificatie van nieuwe, nog niet gekarakteriseerde regulatorische motieven.

Een tweede nadeel van *motiefscreeningmethoden* is dat ze veel vals-positieve motieven identificeren, d.w.z. motieven die geen biologische functie hebben. Het probleem van vals-positieven is nog groter bij

motiefidentificatie in vertebraten, omdat vertebraten gekarakteriseerd worden door lange intergensiche gebieden: de genen van gewervelde organismen worden gescheiden door intergensiche regio's die tot enkele honderden kilobasen lang kunnen zijn. In deze lange sequenties wordt de kans groter een motief toevallig tegen te komen.

De lengte van de vertebrate intergenische gebieden limiteert eveneens de toepassing van *motiefdetectiemethoden*. De lage signaal-op-ruis-verhouding, met name de identificatie van motieven van gemiddeld tussen 5 en 8 bp in sequenties van enkele honderden kb, resulteert in detectie van veel vals-positieven.

Ook de comparatieve aanpak, *phylogenetic footprinting (PF),* is niet ideaal voor de identificatie van regulatorische motieven in vertebrate sequenties: de intergenische gebieden kunnen sterk in lengte verschillen tussen verschillende vertebrate species. Dit is bijvoorbeeld het geval voor kogelvis en zoogdieren. De intergenische gebieden in het kogelvisgenoom zijn gemiddeld 9 keer kleiner dan die in het menselijk genoom. Deze lengteheterogeniteit verhindert vaak een correcte alignering van de orthologe sequenties.

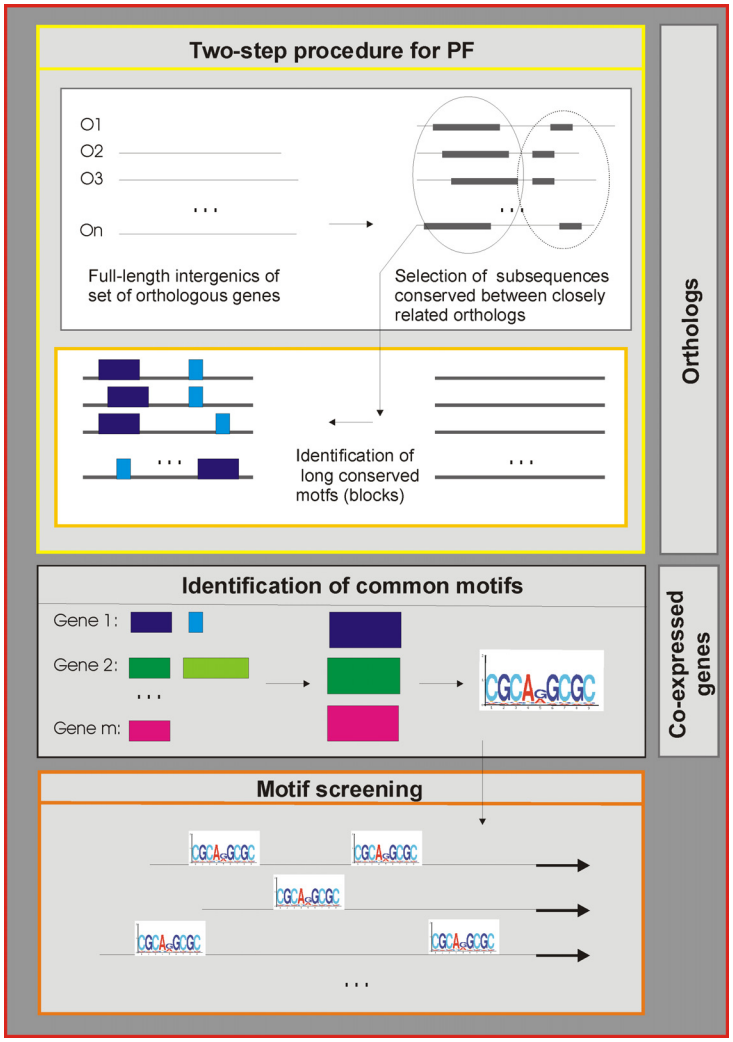
Overzicht van de thesis

Het doel van deze thesis is regulatorische motieven te identificeren in vertebrate organismen. Aangezien geen van de bestaande strategieën op zichzelf optimaal geschikt is voor identificatie van regulatorische motieven in sterk gedivergeerde vertebrate genomen, hebben we in deze thesis de verschillende werkwijzen gecombineerd (zie Figuur N-1). Door verschillende strategieën voor motiefidentificatie te incorporeren, gebruiken we de voordelen van iedere individuele methode. Door elke strategie toe passen op een geschikt data type (bvb. lange vs. korte intergenische regio's; intergenische gebieden van orthologe genen vs. co-gereguleerde genen) minimaliseren we bovendien de nadelen van elke methode. Een concreet voorbeeld: door motiefdetectiemethoden toe te passen op kortere (ge-preselecteerde) sequenties wordt het aantal gedetecteerde vals-positieven gereduceerd. Deze redenering werd gehanteerd voor de ontwikkeling van de tweestapsprocedure, besproken in hoofdstuk 2. Zo wordt in elk hoofdstuk een andere strategie of combinatie van strategieën ontwikkeld en/of gebruikt.

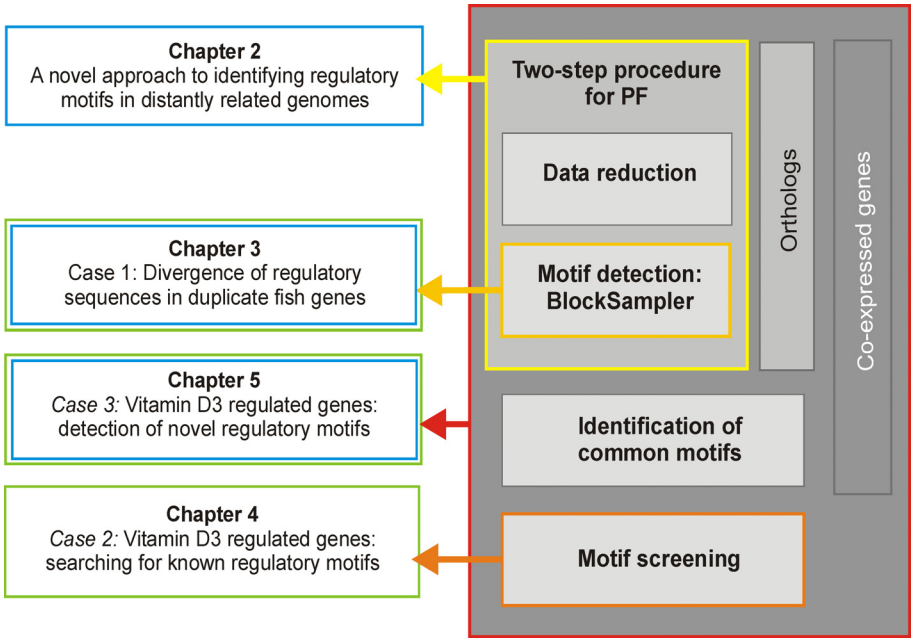
Figuur N-2 geeft voor elk hoofdstuk weer welke strategie(ën) ontwikkeld en/of toegepast werd; in de volgende paragrafen wordt dit in meer detail toegelicht.

Tabel N-1. De oplossingen die worden voorgesteld in deze thesis om de verschillende strategieën voor motiefidentificatie toe te passen op vertebrate organismen. Voor elke strategie worden de belangrijkste beperkingen aangegeven voor de identificatie van regulatorische motieven in vertebraten alsook de oplossingen die werden voorgesteld in deze thesis om deze nadelen te omzeilen.

Strategie	Beperkingen	Oplossingen voorgesteld in deze thesis
Motiefscreening	Beperkt aantal gekende regulatorische motieven Veel vals-positieve motieven (vooral in lange intergenische regio's van vertebrate genen)	*de novo* motief identificatie + motief screening: => identificatie van nieuwe regulatorische motieven
Motiefdetectie	Veel vals-positieve motieven in lange intergenische gebieden van vertebrate genen	Reductie van sequentielengte (datareductie): Vooraf selecteren van niet-coderende sequenties die geconserveerd zijn tussen dichtgerelateerde organismen => toename in signaal-op-ruis-verhouding => minder vals-positieven Motiefdetectie: BlockSampler zoekt naar lange geconserveerde sequentieblokken i.p.v. naar korte motieven => minder vals-positieven
Phylogenetic Footprinting	Lengteheterogeniteit tussen de intergensiche gebieden van verschillende vertebraten bemoeilijkt een correcte alignering met bestaand PF-methoden	Datareductie: Vooraf selecteren van evolutionair geconserveerde intergenische gebieden voor species die gekenmerkt worden door lang intergenische regio's (bvb. zoogdieren), vermindert het verschil in lengte met andere organismen (bvb. kogelvis) Motiefdetectie: Identificatie van motieven met de optimale lengte (lange geconserveerde blokken)



Figuur N-1. Integratie van de verschillende strategieën voor motiefidentificatie. Omdat geen van de bestaande methoden voor de identificatie van regulatorische motieven optimaal is voor motiefdetectie in sterk gedivergeerde vertebraten, hebben we in deze thesis de verschillende werkwijzen gecombineerd. De tweestapsprocedure ('two-step procedure') combineert globale en lokale alignering (*motiefdetectie*) en maakt op die manier identificatie van regulatorische motieven mogelijk in orthologe vertebrate niet-coderende sequenties (*PF*). De combinatie van deze tweestapsprocedure met een clusteringalgoritme resulteerde in de identificatie van betrouwbare motieven in co-gereguleerde genen ('Identification of common motifs'). *Motiefscreening* werd gehanteerd om bijkomende motiefinstanties van de gevonden motieven te lokaliseren ('Motif screening'). Het voorgestelde schema wordt ook gebruikt in Figuur N-2 waar een structureel overzicht gegeven wordt van deze thesis aan de hand van de ontwikkelde en/of toegepaste methodologieën.



Figuur N-2. Overzicht van de thesis. Hoofdstuk 2 behandelt de ontwikkeling en toepassing van de nieuwe tweestapsprocedure voor PF. Hoofdstuk 3 beschrijft de toepassing van de nieuwe implementatie BlockSampler voor de identificatie van regulatorische motieven die subfunctionalizatie ondersteunen. In hoofdstuk 4 wordt gezocht naar gekende regulatorische motieven in de promoterregio van vitamine D_3-gereguleerde genen. Hoofdstuk 5 omvat de ontwikkeling en toepassing van een nieuwe methodologie voor de *de novo* identificatie van regulatorische motieven in co-gereguleerde genen. Deze methode wordt toegepast op een groep van vitamine D_3-gereguleerde genen met hetzelfde expressiepatroon; de geïdentificeerde motieven worden in detail besproken in hoofdstuk 5.

Hoofdstuk 2: Een nieuwe methode voor de identificatie van regulatorische motieven in ver gedivergeerde genomen

Inleiding

Zoals eerder aangehaald, maakt *phylogenetic footprinting* (PF) gebruik van de sequentieconservatie tussen verschillende species om regulatorische motieven af te bakenen. Gesteund door de observatie dat functionele sequentie-elementen trager evolueren dan niet-functionele sequenties, worden potentiële regulatorische motieven geïdentificeerd door geconserveerde sequentie-elementen te detecteren in orthologe intergenische sequenties.

Verschillende algoritmes werden reeds succesvol toegepast voor PF in gist en dichtgerelateerde vertebraten, zoals bijvoorbeeld knaagdieren en zoogdieren. De studie van meer gedivergeerde sequenties draagt echter bij tot een betere afbakening van regulatorische motieven: door orthologe sequenties van minder verwante organismen te vergelijken worden geconserveerde sequentie-elementen scherper afgelijnd tegen de vaak variabele achtergrondsequentie.

Zoals reeds aangehaald in de voorgaande paragrafen zijn de bestaande werkwijzen voor de identificatie van regulatorische motieven niet geschikt voor motiefidentificatie in sterk gedivergeerde vertebraten zoals zoogdieren en vissen. Daarom hebben we een nieuwe procedure ontwikkeld die twee bestaande strategieën combineert: PF en motiefdetectie (zie Figuur N-2). Deze methode bestaat uit twee stappen, vandaar de term tweestapsprocedure, en was specifiek ontwikkeld om regulatorische motieven te identificeren in sterk gedivergeerde vertebrate species.

Tweestapsprocedure voor phylogenetic footprinting

In de eerste stap van de tweestapsprocedure, de datareductiestap, wordt gebruikt gemaakt van een globaal aligneringsalgoritme om regio's te identificeren die bewaard zijn tussen verwante species. Zulke geconserveerde regio's worden namelijk vaak geassocieerd met transcriptiefactor-bindingsplaatsen. Door de intergenische sequenties te beperken tot geconserveerde subsequenties wordt de signaal-op-ruis-verhouding (motief t.o.v achtergrondsequentie) verhoogd voor de tweede stap van de procedure, de motiefdetectiestap. Hiervoor werd een nieuw

probabilistisch algoritme ontwikkeld, BlockSampler, dat gebaseerd is op Gibbs Sampling. BlockSampler gaat op zoek naar lange geconserveerde sequentie-elementen, blokken, i.p.v. korte motieven waardoor het aantal vals-positieven geminimaliseerd wordt. De nieuwe procedure combineert de voordelen van aligneringsmethoden, geschikt voor het vergelijken van lange geconserveerde intergenische sequenties, met de voordelen van motiefdetectiemethoden, die ontwikkeld werden voor het identificeren van kortere geconserveerde regio's (zie Tabel N-1).

Resultaten van tweestapsprocedure op biologische datasets

We hebben de tweestapsprocedure toegepast op vier orthologensets waarvoor functionele, fylogenetisch bewaarde motieven zijn aangetoond *(benchmarkdatasets*) en onze methode identificeerde de meeste vooraf beschreven motieven. Bovendien werden verschillende geconserveerde blokken gedetecteerd die nog niet beschreven waren in de literatuur en ook niet als dusdanig geclassificeerd werden in veel gebruikte databanken (UCSC en UCR). Omdat evolutionair geconserveerde blokken waarschijnlijk verschillende regulatorische motieven bevatten, werd elk van de geïdentificeerde blokken gescreend met de TRANSFAC-databank van gekende vertebrate transcriptiefactor-bindingsplaatsen. Deze screening toonde aan dat de geïdentificeerde blokken een groot aantal homeodomein-bindingsplaatsen bevatten. Naast de gekende TRANSFAC motieven, bevatten de geïdentificeerde blokken ongetwijfeld andere nog ongekende regulatorische motieven. Anderzijds kunnen de blokken eveneens andere biologische functies vervullen waardoor ze geconserveerd zijn gedurende evolutie.

Hoewel het merendeel van de vooraf beschreven regulatorische motieven teruggevonden werden met onze tweestapsprocedure, gingen ook enkele motieven verloren als gevolg van te strenge selectiecriteria. Omdat transcriptiefactor-bindingsplaatsen vaak gegroepeerd voorkomen, veronderstelden we dat de omgeving van een individueel regulatorisch motief (door de aanwezigheid van nabijgelegen regulatorische motieven) eveneens geconserveerd is. Als gevolg worden motieven die in een meer variabele context liggen, met name motieven die geen deel uit maken van een regulatorische module, niet gedetecteerd. Aangezien de tweestapsprocedure generisch is, kunnen de parameters wel aangepast worden.

Tot slot werd de tweestapsprocedure ook toegepast op zes extra datasets, telkens met een verschillende samenstelling (aantal orthologen, aantal niet-zoogdier-orthologen). Hieruit bleek dat de methode algemeen

toepasbaar is voor de identificatie van evolutionair geconserveerde regulatorische motieven in sterk gedivergeerde vertebraten.

Evaluatie van de ontwikkelde tweestapsprocedure

Om de performantie van de tweestapsprocedure te evalueren, vergeleken we de resultaten voor de vier benchmarkdatasets met de resultaten van andere motiefidentificatie-algoritmen: MAVID, TBA, als vertegenwoordigers van meervoudige aligneringsalgoritmen, en MotifSampler, als voorbeeld van een motiefdetectie-algoritme. De tweestapsprocedure is het meest succesvol wanneer de bestudeerde intergenische sequenties te lang worden (typisch voor vertebraten) en/of wanneer het verschil in lengte tussen de orthologe intergenische sequenties te groot wordt.

Vervolgens werd eveneens de bijdrage van elke individuele stap van tweestapsprocedure bekeken. BlockSampler, identificeerde meer vooraf beschreven regulatorische motieven in de benchmarkdatasets in vergelijking met MotifSampler, ongeacht of er een datareductiestap aan vooraf ging. Datareductie, op zijn beurt, leek vooral belangrijk wanneer de bestudeerde intergenische gebieden te lang werden. Dit toont aan dat beide stappen, zowel datareductie als motiefdetectie, essentieel zijn.

Hoofdstuk 3: Divergentie van regulatorische sequenties in geduplicеerde visgenen

Inleiding

Wanneer een gen geduplicеerd wordt, is het mogelijk dat het duplicaat verloren gaat als gevolg van mutaties. In uitzonderlijke gevallen verwerven beide paralogen een nieuwe functie en blijven zo toch in het genoom (neofunctionalizatie). Een derde mogelijkheid is subfunctionalizatie: beide paralogen nemen een (complementair) gedeelte van de oorspronkelijke genfuncties over. Om alle functies van het voorouder gen te verzekeren moeten beide genduplicaten behouden blijven in het genoom. Tot slot kunnen beide kopijen ook bewaard blijven hoewel ze grotendeels dezelfde functies behouden; zulke redundante paralogen bieden het organisme een betere bescherming tegen schadelijke mutaties.

De laatste jaren werd vooral veel aandacht geschonken aan de hypothese van subfunctionalizatie, omdat deze een verklaring biedt voor het grote aantal genen die bewaard zijn gebleven na duplicaties en hun functionele divergentie. Subfunctionalizatie gaat er vanuit dat de functionaliteit van een gen zowel bepaald wordt door de functie van het overeenkomstige proteïne als door het expressiedomein van het gen, d.w.z. waar en wanneer een gen tot expressie komt. Een belangrijke determinant van het expressiedomein van een gen is de transcriptionele regulatie en deze wordt op zijn beurt bepaald door een specifieke combinatie van transcriptiefactor-bindingsplaatsen. Veranderingen in deze transcriptiefactor-bindingsplaatsen kunnen dus een belangrijke antecedent zijn van expressieverschillen en dus ook van sub- en neofunctionalizatie. Er zijn slechts een klein aantal studies die expressieverschillen tussen paralogen in verband brengen met het verschil in regulatorische motieven tussen beide genduplicaten.

Als gevolg van een visspecifieke genoomduplicatie circa 350 miljoen jaar geleden, gevolgd door meer recentere duplicaties, bevatten visgenomen zoals bijvoorbeeld van kogel- en zebravis grote aantallen geduplicеerde genen. Voor verscheidene van deze geduplicеerde genen werd reeds subfunctionalizatie aangetoond. Er bestaat echter geen methode die het mogelijk maakt om op een geautomatiseerde manier regulatorische motieven te identificeren die overeenstemmen met subfunctionalizatie.

Identificatie van motieven indicatief voor subfunctionalizatie

In samenwerking met de onderzoeksgroep 'Bioinformatics & Evolutionary Genomics' van de universiteit Gent, werd onderzocht in hoeverre expressieverschillen tussen paralogen worden gereflecteerd in een verschil van regulatorische motieven tussen beide duplicaten. Meer bepaald, gingen we op zoek naar motieven die aanwezig waren in één paraloog maar afwezig in de andere en die daarmee mogelijk aan de basis liggen van een verschillend expressiedomein. Om dit te onderzoeken ontwikkelden we een generische methode die op basis van *phylogenetic footprinting* regulatorische motieven die differentieel behouden zijn tussen de visparalogen identificeert.

Deze procedure werd toegepast op 12 genfamilies die twee of meer kogelvis- en/of zebravisparalogen bevatten. Voor vijf van deze genfamilies werd het expressieverschil tussen de visparalogen reeds beschreven in de literatuur. In drie van deze '*proof-of-concept*' genfamilies werd minimum één motief geïdentificeerd dat aanwezig was in één van de paralogen en verloren was gegaan in de andere en dat bovendien in overeenstemming was met de experimenteel geobserveerde expressieverschillen. Ook in twee van de overige 7 genfamilies werden verschillen in regulatorische motieven tussen visparalogen genoteerd die mogelijk duiden op subfunctionalizatie.

Om betrouwbare, biologisch functionele motieven te identificeren, werden strenge criteria gehanteerd voor de selectie van motieven. Zo werden bijvoorbeeld enkel motieven in beschouwing genomen die behouden waren over meer dan 450 miljoen jaar evolutie. Door deze maatregel zullen echter ook veel mogelijk functionele motieven verloren gaan, zoals bijvoorbeeld te korte en/of te gedegenereerde motieven en motieven die kenmerkend zijn voor een bepaalde subgroep van de vertebraten. Dit alles resulteert dus waarschijnlijk in een onderschatting van het aantal motieven die subfunctionalizatie ondersteunen. Dit kan verklaren waarom we slechts in drie van de vijf *proof-of-concept*' genfamilies motieven hebben gedetecteerd die de geobserveerde expressieverschillen kunnen verklaren.

Desondanks, de strenge selectiecriteria, zijn we erin geslaagd om in bijna de helft van de bestudeerde genfamilies motieven te detecteren die een (potentieel) verschil in expressiepatroon tussen paralogen kunnen verklaren. Dit toont aan dat subfunctionalizatie waarschijnlijk veel meer voorkomt dan algemeen wordt aangenomen.

Hoofdstuk 6: Besluiten en perspectieven

Besluiten

Omdat geen van de bestaande motiefidentificatiemethoden geschikt waren voor de identificatie van transcriptiefactor-bindingsplaatsen in sterk gedivergeerde vertebrate genomen, hebben wij in deze thesis nieuwe strategieën ontwikkeld die bestaande methoden combineren en die wel met succes toepasbaar zijn op minder verwante vertebrate sequenties.

De belangrijkste methodologische contributies van deze thesis zijn (zie ook Figuur N-1):

- Ontwikkeling van de nieuwe generische tweestapsprocedure voor *phylogenetic footprinting* die regulatorische motieven identificeert in sterk gedivergeerde vertebrate genomen (hoofdstuk 2).
- Ontwikkeling van de eerste generische methodologie voor de identificatie van evolutionair geconserveerde regulatorische motieven die differentieel behouden zijn tussen paralogen en bijgevolg mogelijk duiden op subfunctionalizatie (hoofdstuk 3).
- Ontwikkeling van een generische methodologie voor de identificatie van evolutionair geconserveerde motieven die aanwezig zijn in het intergenisch gebied van meerdere co-gereguleerde genen en dus mogelijk verantwoordelijk zijn voor het gelijkaardige expressiepatroon (hoofdstuk 5).

Deze thesis beschrijft ook de toepassing van de ontwikkelde methodologieën op biologische problemen (zie ook Figuur N-1):

- Studie van subfunctionalisatie
  - Identificatie van regulatorische motieven die overeenstemmen met experimenteel vastgestelde expressieverschillen tussen paralogen (hoofdstuk 3).
- Ontrafelen van het moleculair mechanisme onderliggend aan de werking van vitamine D_3:
  - Detectie van gekende transcriptiefactor-bindingsplaatsen in vitamine D_3-gereguleerde genen toonde aan dat E2F een

cruciale rol speelt in de vitamine D_3-geïnduceerde groei-inhibitie (hoofdstuk 4).

- Identificatie van 31 motieven die aanwezig zijn in het intergenisch gebied van meerdere vitamine D_3-gereguleerde genen en die dus mogelijk betrokken zijn in de moleculaire mechanismen van vitamine D_3 (hoofdstuk 5).

Perspectieven

Hoewel de methodologieën ontwikkeld in deze thesis succesvol zijn voor de identificatie van regulatorische motieven in vertebrate organismen, is er toch nog toekomstig onderzoek mogelijk in dit kader:

- De datareductiestap van de tweestapsprocedure is te rekenintensief om toepasbaar te zijn op grote schaal. Door hiervoor een meer efficiënt algoritme te gebruiken zou de tweestapsprocedure kunnen gehanteerd worden voor de genoomwijde identificatie van evolutionair geconserveerde motieven. Aangezien de bestaande aligneringsmethoden geen foutloze alignering van lange intergensiche sequenties garanderen, is het primordiaal om de toekomstige algoritmen op te volgen en te evalueren.
- Zoals de meeste motiefdetectie-implementaties gaat ook BlockSampler uit van statistisch onafhankelijke invoersequenties, wat niet het geval is voor orthologe sequenties. Daarom wordt er op dit moment gewerkt aan een verbeterd algoritme dat o.a. rekening houdt met de fylogenetische relaties tussen de orthologe sequenties.
- Met de nodige technische verbeteringen zal de tweestapsprocedure genoomwijd kunnen worden toegepast. Dit heeft enkele interessante toepassingen, o.a.:
  - Een overzicht van welke transcriptiefactor-bindingsplaatsen en bijhorende transcriptiefactoren bewaard zijn gebleven gedurende vertebrate evolutie. Dit zou bijdragen tot een beter begrip van welke moleculaire mechanismen universeel zijn in vertebraten.
  - Vergelijking van regulatorische motieven die bewaard zijn in een vertebrate subgroep (bijvoorbeeld zoogdieren) met motieven die bewaard zijn over langere evolutionaire perioden (bijvoorbeeld zoogdieren en vissen). Op die manier kunnen we nagaan welke regulatorische mechanismen

bepalen wie we zijn en welke mechanismen een rol spelen in een bredere populatie van organismen (hier zijn verschillende niveaus mogelijk: mens t.o.v. zoogdieren, zoogdieren t.o.v. vertebraten, enzovoort).

- Ook de *de novo* motiefdetectie zoals beschreven in hoofdstuk 5 kan worden geoptimaliseerd. Op dit moment worden de motieven geïdentificeerd door sequentieel fylogenetische informatie (orthologen) en co-expressie te gebruiken. Er wordt gewerkt aan een meer geavanceerd algoritme waarin simultaan gebruik wordt gemaakt van beide informatiebronnen.
- De vergelijking van meerdere gemiddeld tot sterk gedivergeerde organismen bevordert de identificatie van regulatorische motieven d.m.v. *phylogenetic footprinting*: door sequenties te aligneren die evolutionaire veranderingen hebben ondergaan, worden geconserveerde functionele sequentie-elementen, zoals transcriptiefactor-bindingsplaatsen, scherper afgelijnd tegenover het niet-geconserveerde functieloze DNA. De sterktes van onze tweestapsprocedure, die de vergelijking van een aantal sterk gedivergeerde sequenties mogelijk maakt, zullen dus steeds duidelijker worden naarmate meer en meer vertebrate genomen publiek beschikbaar worden.
- Een direct gevolg van de ontwikkelde *phylogenetic footprinting* methode is de mogelijkheid om nieuwe transcriptiefactor-bindingsplaatsen te ontdekken die kunnen bijdragen tot de opheldering van totnogtoe ongekende reactiemechanismen. Het feit dat de tweestapsprocedure, in tegenstelling tot de meeste andere motiefidentificatiemethoden, regulatorische motieven identificeert in lange intergenische sequenties (bvb. ver stroomopwaarts van een gen) impliceert dat deze methode meer motieven zal terugvinden in vergelijking met methoden die zoeken in kortere sequenties.
- Een zeer interessante toepassing is de genoomwijde subfunctionalizatiestudie. Door de ontwikkelde methodologie toe te passen op alle paralogen van een organisme, kunnen we een inschatting maken van het aantal genen die van functie zijn veranderd gedurende evolutie. Vissen vormen zeer boeiend studiemateriaal: de grote diversiteit aan vissoorten (~ evolutionair succes) wordt vaak toegeschreven aan het groot aantal genen aanwezig in visgenomen, die het gevolg zijn van opeenvolgende duplicaties. Door het aantal veranderingen in genfunctie binnen een visgenoom te kwantificeren zouden we deze hypothese kunnen evalueren.

- De methoden ontwikkeld in deze thesis kunnen gebruikt worden om alle potentiële doelwitgenen van elk mogelijk regulator protein te bepalen. Met behulp van *phylogenetic footprinting* identificeren we eerst regulatorische motieven. Deze motieven kunnen gebruikt worden om genoomwijd alle intergenische gebieden te screenen. Op die manier genereren we een compendium met al de genen die mogelijk gereguleerd worden door dezelfde regulator. In het geval van een gekend motief en/of regulator draagt zo een compendium bij tot het ontrafelen van de werking van dit regulatorisch protein. Een andere mogelijkheid is dat we vertrekken van een ongekend motief; in dit geval kan een lijst met alle potentiële doelwitgenen gebruikt worden om de functie van het motief en zijn regulator af te leiden afgaande op de functies van de doelwitgenen.
- Alle methoden die in deze thesis ontwikkeld en/of toegepast werden zijn generisch. De toepassingen zijn dus eindeloos:
  - Zoals reeds vermeld werd zou genoomwijde toepassing van de verschillende strategieën bijdragen tot de identificatie van vele (nieuwe) transcriptiefactor-bindingsplaatsen. Dit kan leiden tot een beter inzicht in de regulatorische mechanismen en een beter begrip van hoe vertebrate organismen geëvolueerd zijn en hoe ze functioneren.
  - Naast grootschalige toepassingen, kunnen de methodologieën ook gebruikt worden om één specifiek moleculair mechanisme of reactieweg (zoals bijvoorbeeld vitamine D_3-geinduceerde groei-inhibitie) te ontrafelen. Dit zou bijvoorbeeld bijdragen tot doorgronden van de onderliggende mechanismen van autoimmuunziekten, kanker, …

Publications

- Verlinden L, Eelen G, Beullens I, Van Camp M, Van Hummelen P, Engelen K, **Van Hellemont R**, Marchal K, De Moor B, Foijer F, Te Riele H, Beullens M, Bollen M, Mathieu C, Bouillon R and Verstuyf A. Characterization of the condensin component Cnap1 and protein kinase Melk as novel E2F target genes down-regulated by 1,25-dihydroxyvitamin D_3. 2005. The Journal of Biological Chemistry, 280 (45): 37319-37330.
- **Van Hellemont R**, Monsieurs P, Thijs G, De Moor B, Van de Peer Y and Marchal K. A novel approach to identifying regulatory motifs in distantly related genomes. Genome Biology, 2005, 6 (13): R113.1-R113.18.
- Verlinden L, Eelen G, **Van Hellemont R**, Engelen K, Beullens I, Van Camp M, Marchal K, Mathieu C, Bouillon R and Verstuyf A. 1alpha,25-Dihydroxyvitamin D(3)-induced down-regulation of the checkpoint proteins, Chk1 and Claspin, is mediated by the pocket proteins p107 and p130. The Journal of Steroid Biochemistry and Molecular Biology, 2006, in press.
- **Van Hellemont R**, Blomme T, Van de Peer Y and Marchal K. Divergence of regulatory sequences in duplicated fish genes. Invited book chapter in 'Genome Dynamics vol. 3: Gene and Protein Evolution', 2007 (in press). Editor: Volff, J.-N.

Contents

Chapter 1

Introduction

1.1 Context of the thesis

Since the publication of the first draft of the human genome in 2001 (Lander et al. 2001; International Human Genome Sequencing Consortium. 2001; Venter et al. 2001), many higher eukaryotic genomes have been sequenced. In February 2007, the Genomes OnLine (GOLD) database contains 514 complete genomes including 47 eukaryotic genomes (Liolios et al. 2006); among these are 7 chordates: *Danio rerio*, *Fugu (Takifugu) rubripes* (Aparicio et al. 2002), *Homo sapiens* (International Human Genome Sequencing Consortium. 2001), *Mus musculus* (Mouse Genome Sequencing Consortium 2002), *Rattus norvegicus* (Rat Genome Sequencing Project Consortium 2004), *Tetraodon nigroviridis* (Jaillon et al. 2004) and the 'primitive' chordate *Ciona intestinalis*. (Dehal et al. 2002). Furthermore, there are hundreds of ongoing genome sequencing projects including amphibians, e.g., frog (*Xenopus laevis, Xenopus tropicalis*); birds, e.g., *chicken (Gallus gallus*, International Chicken Genome Sequencing Consortium 2004); fish, e.g., salmon (*Salmo salar*) and mammals, e.g., cat (*Felix catus*), chimp (*Pan troglodytes*, The Chimpanzee Sequencing and Analysis Consortium 2005), dog (*Canis familiaris,* Kirkness et al. 2003), horse (*Equus caballus*) and pig (*Sus scrofa*); and so on. Recently, Green et al. (2006) sequenced part of the Neanderthal DNA. This source of DNA is, among others, interesting from a palaeontologic point-of-view, for example, to calculate human-Neanderthal divergence time (Noonan et al. 2006).

This increasing amount of genome information is stored and made available for scientists through genome databases such as Ensembl (Birney et al. 2006) , National Centre for Biotechnology Information (NCBI), NIH Intramural Sequencing Center (NISC) and the Gene Index Project (TGI) (Figure 1-1). As new genome sequences are being finished at a high rate, the most important public sequence databases are doubling in size every 18 months (Stahler et al. 2006). Such a dynamic development reminds us of

Moore's law concerning information technology; this law described –in its most popular formulation- the doubling of the number of transistors on integrated circuits (a rough measure of computer processing power) every 18 months.

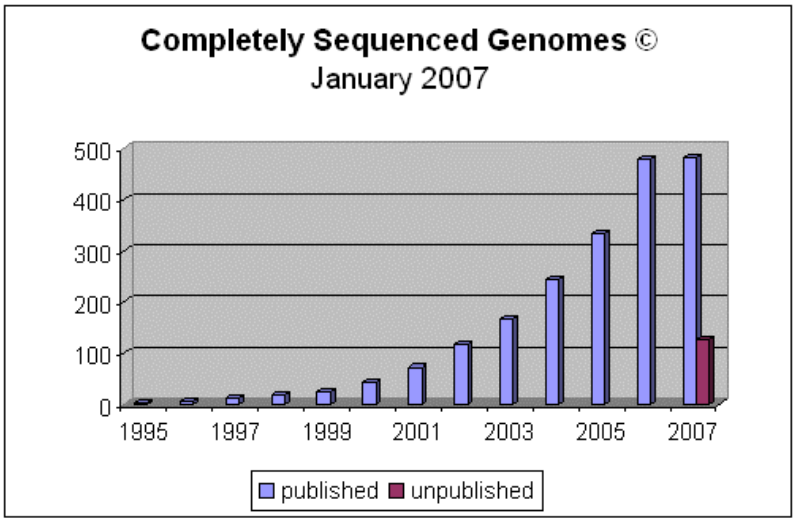


Figure 1-1. The growth of the GOLD database. GOLD is a web resource for comprehensive access to information regarding complete and ongoing genome sequencing projects (Liolios et al. 2006). This graph shows for each year indicated the amount of completely sequenced genomes at that time. The graph shows an exponential growth in genomic data available. From http://www.genomesonline.org.

These complete genome sequences provide the basics to understanding the complexity of entire organisms. Furthermore, complete genomes can be used for comparative genomics: by comparing sequences of distinct organisms we look for sequence elements that are conserved among evolutionary time. Such conserved motifs are probably functional elements such as protein-coding genes and exons (Bafna and Huson 2000; Batzoglou et al. 2000; Wiehe et al. 2001; Korf et al. 2001; Novichkov et al. 2001; Rinner and Morgenstern 2002; Morgenstern et al. 2002; Meyer and Durbin 2002; Parra et al. 2003; Blayo et al. 2003; Taher et al. 2004; Haubold and Wiehe 2004; Meyer and Durbin 2004). The underlying rationale is that a random mutation in a functional region is usually deleterious to the organism, and therefore unlikely to accumulate over time (Wiehe et al. 2000). Each new genome that is sequenced will further resolve those regions that are of critical functional importance (Kirkness et al. 2003).

When comparing the genomic sequence of human and mouse, 5% of their genomes seem to be under purifying selection. Protein-coding regions comprise circa 1.5% of the human genome and can thus not solely account for this sequence conservation. Together with the observation that the

intergenic regions between organisms are more conserved than expected for functionless 'junk' DNA -for instance, 4% for dog, human and mouse (Kirkness et al. 2003)-, this implies that the mammalian genome contains many additional functional conserved elements, such as untranslated regions, non-protein-coding genes, regulatory motifs, etc. (Dermitzakis et al. 2002; Mouse Genome Sequencing Consortium 2002; Chiaromonte et al. 2003).

In this thesis we focused on identifying regulatory motifs based on their conservation profile. This chapter first gives a brief introduction to transcription regulation of eukaryotic genes (§1.2). Secondly, we define what is understood by a 'regulatory motif' and how it functions in the machinery that drives gene expression in eukaryotic genes (§1.3). After giving a concise overview of the distinct methods that exist to identify regulatory motifs (§1.4), we end this introductory chapter with an outline of the thesis (§1.5).

1.2 Transcription regulation in eukaryotes

Transcription is the process during which genetic information is transcribed from DNA to RNA. The 'expression' of a gene designates the level of messenger RNA (mRNA) present in the cell transcribed from that gene. For most protein-coding genes the level of expression varies along with the circumstances, i.e., developmental stage, cell type, nutrient level, etc. The expression level of each individual gene is mostly controlled at the level of transcription (Wray et al. 2003). Transcription regulation is a highly dynamic process that involves a combination of factors: the general transcription initiation factors that make up the basal transcription apparatus (§1.2.1), sequence-specific DNA binding factors that bind to up- or downstream regulatory elements (§1.3) and associated accessory factors.

1.2.1 Basal transcription process

Eukaryotic protein-coding genes are transcribed by the RNA polymerase II (RNAPII) holoenzyme complex (Lee and Young 2000). This complex consists of RNAPII and a set of basal transcription factors (TFs), namely TFIIA, B, D, E, F and H, as illustrated Figure 1-2 (Tjian 1996).

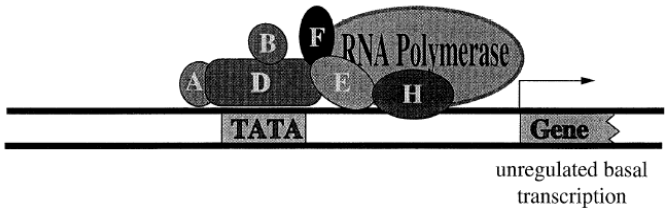


Figure 1-2. The RNA polymerase II holoenzyme. The picture depicts the collection of basal factors (TFIIA, B, D (or TBP), E, F and H) that assemble along with RNA polymerase II to form the RNAPII holoenzyme. From Tjian (1996).

Assembly of the RNAPII holoenzyme complex on the basal promoter initiates transcription (Figure 1-3). Although basal promoter sequences differ among genes, for many genes the critical binding site is the TATA box, usually located circa 25-30 bp upstream of the transcription start site (tss). In such promoters, the attachment of the TATA-binding protein (TBP, also known as TFIID) to the TATA box is a crucial step in transcription initiation (Figure 1-2, Figure 1-3, Kuras and Struhl 1999). Some genes, however, contain an initiator element instead of the TATA box or neither of both. In these cases, TBP binds to the DNA in a sequence-independent manner; proteins that bind to other motifs in the basal promoter facilitate this (Wray et al. 2003). Once TBP attaches to the DNA, several TBP-associated factors (TAFs) guide the RNAPII holoenzyme complex to DNA. Transcription factors binding at other sites can modulate this attachment in positive or negative way (Lee and Young 2000; Lemon and Tjian 2000). After the RNAPII holoenzyme complex assembles onto the DNA a second contact is established at the tss (Wray et al. 2003).

By itself a basal promoter initiates transcription at a very low rate (Wray et al. 2003). Moreover, the transcription initiation factors binding to the basal promoter and assisting the initiation of transcription are omnipresent, providing little regulatory specificity (Kuras and Struhl 1999; Lemon and Tjian 2000). Producing functionally significant levels of mRNA requires the sequence specific binding of transcription factors (TFs) to DNA motifs, i.e., transcription factor binding sites (TFBSs), outside the basal promoter (Lemon and Tjian 2000). These TFBSs, also regulatory motifs, are discussed in more detail in §1.3.

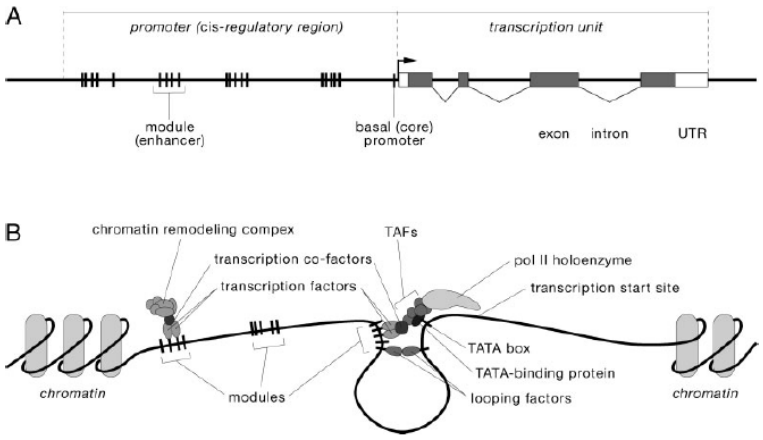


Figure 1-3. Promoter structure and function. (A) Organization of a generalized eukaryotic gene, showing the relative position of the transcription unit -consisting of exons, introns and untranslated regions (UTR)-, basal promoter (black box with bent arrow) and transcription factor binding sites (vertical bars) (B) Idealized promoter in operation. Initiating transcription requires multiple different proteins, which interact with each other in specific ways. These include the RNAPII holoenzyme complex, TATA-binding protein (TBP), TAFs (TBP-associated factors), transcription factors (TFs), transcription cofactors and chromatin remodelling complexes. For further details we refer to the main text. From Wray et al. (2003).

1.3 Regulatory elements

The observation that eukaryotic genomes -from nematodes to insects to mammals- contain a similar number of protein-coding genes, lead to hypothesis that the organismal complexity lies in, among others, transcription factor binding sites (TFBS) and transcription factors (TFs) (Rodriguez-Trelles et al. 2003; GuhaThakurta 2006). This hypothesis is supported by the findings of Nelson et al. (2004): they found that developmentally important genes are flanked by significantly more non-coding sequences than genes with less complex functions (e.g., housekeeping genes) and thus are likely to contain many more regulatory motifs. Furthermore, they showed that such regions of high regulatory complexity are significantly longer in more complex organisms (Nelson et al. 2004).

The binding of a TF to a specific TFBS, also referred to as regulatory motif, is one of the key factors in regulation of gene expression. (§1.2 and §1.3.1). As explained in §1.3.2 regulatory motifs are often grouped in clusters, regulatory modules. These modules are the building blocks of gene regulatory networks (see §1.3.3).

1.3.1 Regulatory motifs

1.3.1.1 Definition

TFBSs or regulatory motifs are stretches of DNA that are recognized sequence-specifically by a TF that is required to control the expression of the target gene (see Figure 1-5); this TF can be an activator, enhancing the transcription of the target gene, or a repressor doing the opposite. Regulatory motifs specify and anchor the TFs in appropriate positions with respect to one another and to the basal transcription apparatus (see §1.2.1). These TFs, and other proteins that in turn bind to them, determine the rate of transcription and mediate the accurate activation and repression of the gene in developmental time and morphological space (Arnone and Davidson 1997).

Most regulatory motifs are 5 to 8 nt long. Their presence is most often associated with the promoter region of the gene (i.e., the intergenic region located immediately upstream of the start of the gene), but recently it has been shown that they also occur at long distances upstream from the gene they target (e.g., *pax6* in mouse: Kammandel et al. 1999; Woolfe et al. 2005, see Figure 1-3 A). Furthermore, regulatory motifs sometimes occur in the untranslated regions (UTRs; e.g., *scr* in Drosophila melanogaster: Calhoun et al. 2002), the introns (e.g., *ccr5* in humans: Bamshad et al. 2002) downstream (3') of the transcription unit (*bmp5* in mouse: DiLeone et al. 1998) and, rarely, within a coding exon (keratin 18 in humans: Neznanov et al.1997). This diversity of positions is possible because DNA looping allows interaction between proteins associated with DNA and distant binding sites (Wray et al. 2003).

Known TFBS are made publicly available through databases. Examples of such databases are EPD (Praz et al. 2002), JASPAR (Sandelin et al. 2004; Vlieghe et al. 2006), PlantCARE (Lescot et al. 2002) and TRANSFAC (Wingender et al. 2001; Matys et al. 2003). Little is known about the amount of regulatory motifs present in mammalian genomes, but the number of such motifs is expected to be an order of magnitude higher than the number of protein-coding genes, i.e., in the order of hundreds of thousands or more (Yuh et al. 1998). The widely used TRANSFAC database (professional release 10.2, June 2006) contains 584 models for vertebrate TFBSs (Wingender et al. 1996; Wingender et al. 2001; Matys et al. 2003). This shows that our current knowledge of these DNA binding sites is severely limited (GuhaThakurta 2006). Although many methods have been developed to identify regulatory motifs, much more research is needed. An overview of the main existing methods is given in §1.4.

1.3.1.2 Representation

Although the binding of a TF to a TFBS is sequence-specific, most TFs allow some variability in their binding site, i.e., most TFBSs can tolerate a few nucleotide substitutions without losing their functionality (Latchman 1998). There are several ways of representing a regulatory motif. In this paragraph we consider three frequently used representations: the consensus sequence (§1.3.1.2.1), the matrix model (§1.3.1.2.2) and the motif logo (§1.3.1.2.3).

1.3.1.2.1 Consensus sequence

The consensus sequence is widely used to represent regulatory motifs, however, the definition is rather arbitrary. In general the consensus sequence matches all of the binding sites closely, but not necessarily exactly (Stormo 2000); it consists of the most frequent nucleotide at each position (Pavesi et al. 2004a) (Figure 1-4 B). The alphabet used for the consensus sequences is the IUPAC degenerate alphabet (see Table 1-1; Nomenclature Committee for the International Union of Biochemistry 1986). The consensus representation is used in chapters 2 and 3.

Table 1-1. IUPAC codes used to denote ambiguous positions in nucleotide sequences. From Pavesi et al. 2004.

IUPAC	Nucleotides	Mnemonics
A		Adenine
C		Cytosine
G		Guanine
T		Thymine
R	A or G	Purines
Y	C or T	Pyrimidines
W	A or T	Weak hydrogen bonding
S	G or C	Strong hydrogen bonding
M	A or C	Amino group at common position
K	G or T	Ketogroup at common position
H	A, C, T	Not G
B	C, G, T	Not A
V	A, C, G	Not T
D	A, G, T	Not C
N	A, C. G. T	Any

1.3.1.2.2 Matrix model

A more accurate alternative for TFBS representation is the matrix model. Like the consensus sequence this representation is also derived from the alignment of a set of binding sites for a particular TF. In its most simple form, the count matrix, this matrix lists the number of occurrences of each nucleotide at each position of an alignment (Figure 1-4 B).

From the count matrix, a position specific frequency matrix (PSFM) can be constructed by calculating the frequencies of each nucleotide at each position and by introducing pseudocounts. These pseudocounts are very small frequencies that are introduced where a zero occurs in the PSFM: a zero indicates that this particular nucleotide is not observed at this particular position in the binding sites instances aligned, but this does not guarantee that this does not occur anywhere in the genome. The PSFM is used in the MotifLocator, MotifSampler, MotifScanner and BlockSampler algorithms (see chapters 2, 3, 4 and 5).

An alternative matrix model is the position weight matrix (PWM) (Figure 1-4 C). In this motif model weights are calculated using the following formula (Hertz and Stormo 1999; Stormo 2000):

$$w_{ij} = \ln \frac{(n_{i,j} + p_i) / (N + 1)}{b_i} \approx \ln \frac{f_{i,j}}{b_i} \qquad [1.1]$$

where $n_{i,j}$ is the total number of occurrences of nucleotide i at position j, p_i is the pseudocount for nucleotide i, N is the total number of sequences, b_i is the single nucleotide frequency in the genome for nucleotide i and $f_{i,j}$ is the frequency of nucleotide i at position j.

1.3.1.2.3 Motif logo

A motif logo graphically represents the amount of information that is stored at each position of the binding site (Figure 1-4 D). This type of representation will be used in chapter 5.

The information content at specific position is calculated as follows (Schneider et al. 1986):

$$I_i = \sum_{b=A}^{T} f_{b,i} \log_2 \frac{f_{b,i}}{p_b} \qquad [1.2]$$

where i is the position within the site, b refers to the different nucleotides, $f_{b,i}$ is the observed frequency of each nucleotide at position i and p_b is the frequency of base b in the whole genome. The information content, I_i, is 0 for positions where the frequency of each nucleotide at that position is equal (0.25), given that the background frequency p_b is also 0.25. For positions that

are completely conserved for one nucleotide I_i equals 2. I_i is also known as the relative entropy or the Kullback-Leibler distance.

I_i is used as total height of the stack in the motif logo (Figure 1-4 D): i.e., the sum of the heights of each nucleotide corresponds to the information that is stored at that position.

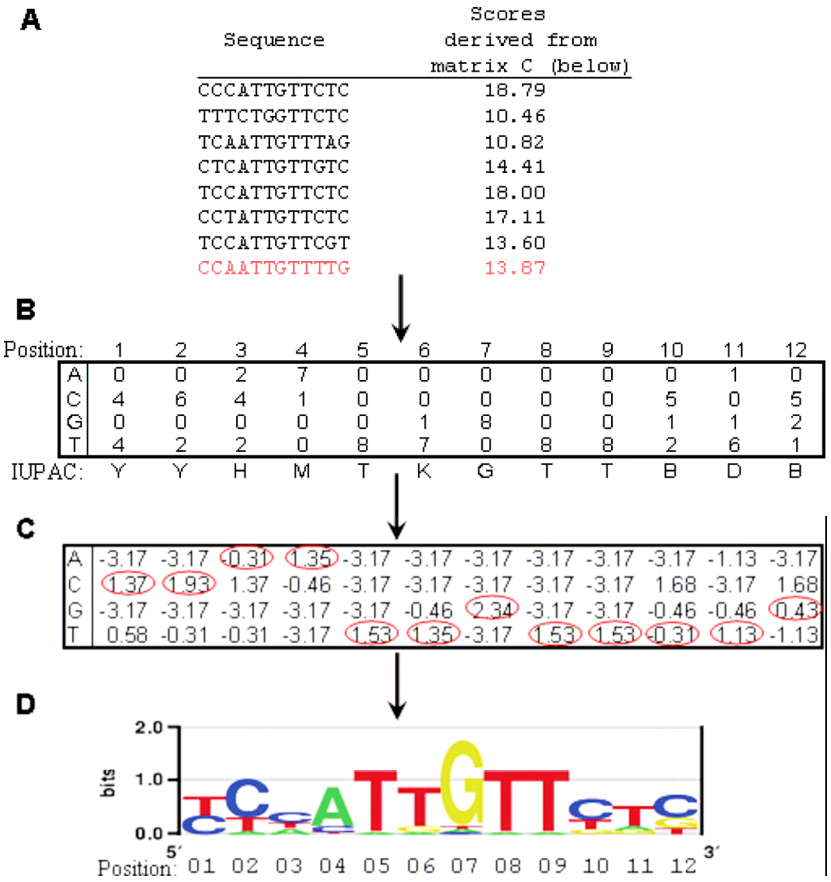


Figure 1-4. Different representations of transcription factor binding sites. (A) the collection of known binding sites, in this case Rox-1 binding sites taken from the SCPD database (Zhu and Zhang 1999). (B) Count matrix and consensus sequence of the eight binding sites from panel A. The cells represent the number of times a base *i* is observed at position *j* in the alignment of sites. (C) Position weight matrix (PWM). A pseudocount of 1 was added to the alignment before deriving the weights. (D) Motif logo representation of the alignments, visually showing the information content and conservation at each of the alignment positions. From GuhaThakurta 2006.

1.3.2 Regulatory modules

A regulatory module is a group of multiple regulatory motifs (Figure 1-5). Such modules vary in length from a few hundreds base pairs (bp) (e.g., Balmer and Blomhoff 2006) to more then 100 kilo basepairs (kb) and typically contain 6 to 15 binding sites recognized by 4 to 8 different TFs (Arnone and Davidson 1997; Balmer and Blomhoff 2006).

A regulatory module contains several TFBSs that contribute in various ways to the regulatory output: it can transmit regulatory outputs to the basal transcription apparatus and/or to other regulatory modules. Often the constituting regulatory motifs are recognized and bound by different type of TFs with diverse functions, such as factors controlling relations between the module and the basal transcription apparatus or other modules, factors directly activating or repressing the target gene, … (Arnone and Davidson 1997). Together with the genes they regulate and isolated TFBSs, regulatory modules are the building stones of gene regulatory networks (see §1.3.3; Wray et al. 2003).

1.3.3 Gene regulatory network

The linkage between regulatory systems (motifs and modules) together with the genes that they govern is called a gene regulatory network (GRN) (Arnone and Davidson 1997). The deciphering of regulatory networks underlying various developmental/biological processes is a major challenge of the post-genomic period (Davidson et al. 2002).

An example of a rudimentary GRN is illustrated in Figure 1-5: Due to an external stimulus (starvation, developmental stage, signals from adjacent cells, …) the transcription of gene 1 is (directly or indirectly) activated through binding of the orange transcription activating complex on the orange module. Gene 1 encodes a TF (protein 1) that recognizes a regulatory motif (i.e., part of the blue regulatory module) controlling expression of gene 2. The gene expression of this second gene results in a (augmented) production of protein 2. This protein can, for instance, be another TF inducing or repressing expression of downstream genes; these genes then also belong to the example GRN. Alternatively, protein 2 can be an effector-protein, e.g., a transporter protein.

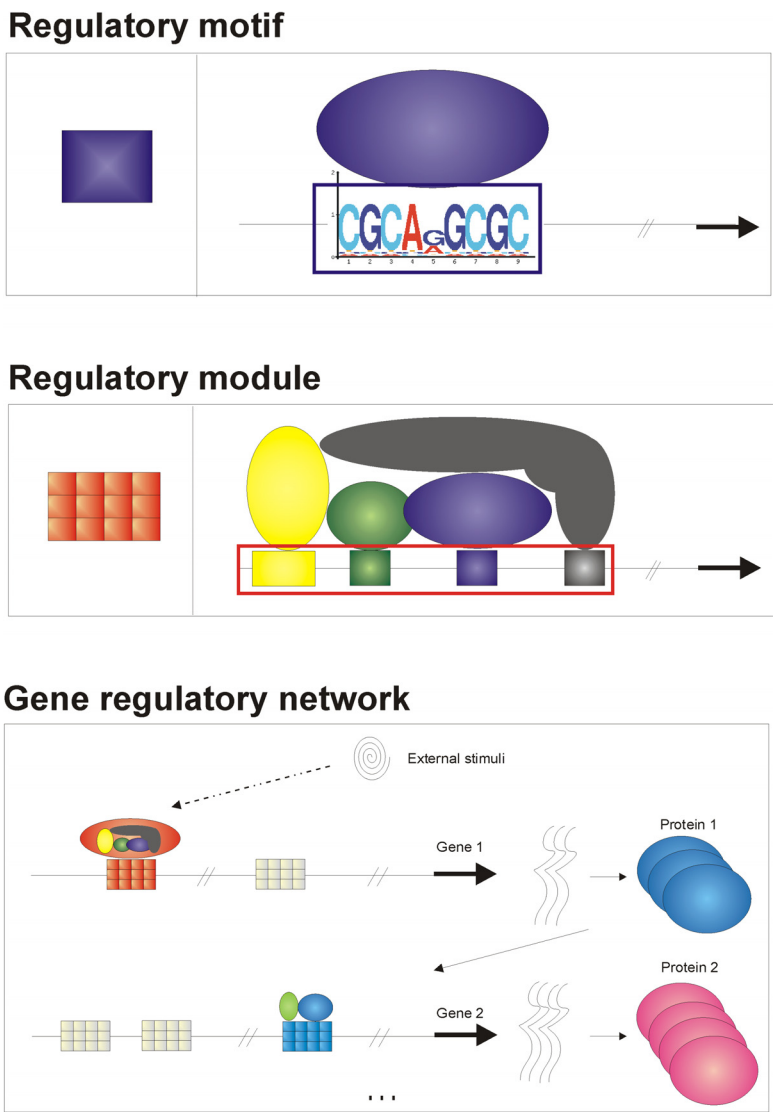


Figure 1-5. From regulatory motif to gene regulatory network. The top diagram represents the sequence-specific binding of a transcription factor (TF) to a transcription factor binding site (TFBS) or regulatory motif. This motif is represented by its consensus sequence (see §1.3.1.2.1). The middle diagram illustrates the binding of several TFs to the regulatory motifs of a regulatory module. In this specific case the TFs interact with each other to form a complex, but this is not necessary. The lower diagram depicts a very rudimentary gene regulatory network (GRN): for a more detailed description we refer to the text (see §1.3.3).

1.4 Computational discovery of regulatory motifs

Because empirical validation of binding sites is laborious, several computational methods have been developed for identification of regulatory motifs. In this section we will discuss a few of the many approaches that have been developed. Rather than exhaustively covering the literature, we aim at capturing the key concepts. For a more thorough description of the existing motif discovery tools we refer to some interesting reviews by Bulyk (2003), Ureta-Vidal et al. (2003), Wasserman and Sandelin (2004) and Sandve and Drablos (2006) as well as a comparative studies of several tools by Pollard et al. (2003) and Tompa et al. (2005).

A first group of methods uses databases of known TFBSs to scan sequences for potential TFBSs, motif-screening methods. These are discussed in §1.4.1. Alternative approaches look for regulatory motifs without relying on known TFBSs; these are discussed in §1.4.2.

1.4.1 Motif screening

A first category of methods was developed to identify known TFBSs, i.e., motif screening. The identification of potential regulatory motifs starts with matching a chosen motif model against the input sequence(s). This matching can be either string- or matrix-based. Once the potential sites are identified, some statistic is applied to identify the significant motif instances.

1.4.1.1 Detection of individual motifs

1.4.1.1.1 String-based screening

A first approach to find potential regulatory motifs is to perform a string search with a known TFBS. Instances of the motif are identified as subsequences within the sequence that almost exactly match the known binding site. Scores are calculated by counting the number of mismatches between the potential and the known TFBS. A threshold on the score is used to select the relevant, best-scoring motif instances.

1.4.1.1.2 Matrix-based screening

Alternatively, regulatory motifs can be discovered by scoring sequences with a matrix model of a known TFBS (see §1.3.1.2.2). Such motif models can be derived from databases, such as TRANSFAC (Wingender et al. 2001) and JASPAR (Sandelin et al. 2004) (see §1.3.1.1).

This method involves adding the matrix weights of each occurring letter in the searched sequence together and normalizing for the length of the matrix. The score ($W'(x)$) is then normalized with respect to the minimal (W_{min}) and maximal (W_{max}) possible score:

$$W'(x) = \frac{W(x) - W_{min}}{W_{max} - W_{min}} \qquad [1.3]$$

where $W(x)$ is the score for a given subsequence x, W_{min} is the sum of the smallest weights at each position and W_{max} is the sum of the highest weights at each position. The normalized score varies between 0 and 1.

In order to decide whether a certain subsequence is a putative TFBS a threshold for the normalized score is most commonly used. Examples of implementations that use such matrix scoring method are MatInd and MatInspector (Quandt et al. 1995), MATRIX SEARCH (Chen et al. 1995), SIGNAL SCAN (Prestridge 1991; Prestridge 1996), rVISTA (Loots et al. 2002) and TFM-Explorer (Defrance and Touzet 2006). Also the in-house developed algorithm MotifLocator was developed to score sequences with PSFMs on a score cut-off basis (Thijs et al. 2002b; Coessens et al. 2003).

A more sophisticated way of selecting motif instances is implemented in the MotifScanner algorithm, also developed at ESAT-SCD (Thijs et al. 2002b; Coessens et al. 2003). MotifScanner starts from the probabilistic sequence model based on the assumption that motifs are hidden in a noisy background sequence. This sequence model is used to estimate the number of instances of a motif model in a sequence.

The multitude of TFs in eukaryotic genomes (mammalian genomes are estimated to have circa 2000 TFs (International Human Genome Sequencing Consortium. 2001; Mouse Genome Sequencing Consortium 2002)) together with the observation that TFs bind to different binding sites that are often short and imprecise, implies that every kilo base (kb) of genomic DNA contains dozens of potential TFBSs on the basis of random similarity (Carroll et al. 2001; Stone and Wray 2001; Wray et al. 2003). Therefore, many of the potential binding sites identified with the programs described in this paragraph have no biological function and are simply spurious matches to known TFBSs.

In order to augment the confidence in the regulatory motifs identified some variants have been developed to the classical motifs screening schemes. CONREAL, for instance, that combines motif screening with sequence conservation (Berezikov et al. 2004; Berezikov et al. 2005). This algorithm is based on the observation that, unlike functionless background sequences, functional elements, including regulatory motifs, tend to be conserved during evolution. This principle forms the basis of phylogenetic footprinting, which is explained further on, in §1.4.2.2.

1.4.1.2 Detection of regulatory modules

Besides discovery of individual TFBSs, it is interesting to assess whether we can identify regulatory modules, i.e., groups of regulatory motifs that occur near each other (see §1.3.2). Such modules possibly represent TFBSs for cooperatively acting TFs. Several approaches have been developed to identify regulatory modules.

A first category of methods uses the 'module scanning' approach. These algorithms detect a joint occurrence of a known combination of known TFBSs within a certain sequence window. Most of the module scanners use the 'sliding window' approach: they search for a subsequence that contains motif instances for all or most of the known TFBSs. These methods do not take into account the order the motif instances occur in or the spacing between the different motif instances. MSCAN, for instance, calculates the combined statistical significance of the motif instances of a given set of PWMs in a window (Johansson et al. 2003; Alkema et al. 2004). Other examples of such algorithms are Co-Bind (GuhaThakurta and Stormo 2001), which utilizes a Gibbs sampling strategy to model the cooperativity between two transcription factors and defines position weight matrives for the binding sites, ModuleScanner (Aerts et al. 2003) and ModuleFinder (Philippakis et al. 2005). Other module scanners use a hidden markov model implementation that takes distance constrains between TFBSs into account. Examples of such algorithms are Cluster-Buster (Frith et al. 2003), MCAST (Bailey and Noble 2003) and Stubb (Sinha et al. 2003).

Alternatively, it is interesting to search for a joint occurrence of a new combination of known TFBSs. ModuleSearcher , developed at ESAT-SCD was developed to identify such modules based on overrepresentation in a set of co-regulated genes (Aerts et al. 2003; Aerts et al. 2004). CoMoDis, is an alternative approach, that discovers regulatory modules by starting from a single known regulatory motif that is thought to be important in the regulation of a set of co-regulated genes (Donaldson and Gottgens 2006).

1.4.2 *De novo* identification of regulatory motifs

De novo motif identification methods aim at identifying regulatory motifs without any prior knowledge about which specific binding site is searched for. These categories of algorithms include methods starting from co-expressed genes (§1.4.2.1) and methods that make use of comparative genomics, i.e., phylogenetic footprinting (§1.4.2.2). Finally, for the sake of completeness, we quickly mention a few algorithms that have been developed to identify regulatory modules in a *de novo* manner (§1.4.2.3).

1.4.2.1 Motif detection algorithms

The first category of *de novo* procedures aims at identifying overrepresented motifs in the intergenic regions of co-expressed genes. These algorithms are further referred to as motif detection algorithms (for instance, see chapter 2). The rationale is that genes showing a similar expression pattern (under specific conditions) are likely to share the same transcriptional regulation mechanism, hence the same regulatory motifs. This class of algorithms can be divided in two groups: probabilistic (§1.4.2.1.1) and combinatorial methods (§1.4.2.1.2). It is not our goal to give a detailed description of the different algorithms; we only want to indicate the basic principles.

1.4.2.1.1 Probabilistic methods

The probabilistic methods use a PSFM to represent regulatory motifs (see §1.3.1.2.2). Moreover, they all start from a probabilistic sequence model: except for the motif instances, the sequences are considered as noisy sequences where each base is generated according to a specific background model.

The oldest probabilistic motif detection tools use the Expectation-Maximization (EM) method. Concerning motif finding, EM assumes that each sequence contains exactly one motif instance, from which the starting position is unknown (missing value from the data) (Hertz et al. 1990). In order to determine the starting positions, each subsequence is scored with the current motif model. These updated probabilities are used to re-estimate the motif model. This procedure is repeated until convergence. The main drawback of EM is its sensibility to the initialization point; convergence to the global maximum is not guaranteed. Moreover, the assumption that each sequence carries exactly one motif instance is not biological correct. Bailey and Elkan (1995) developed a more advanced EM implementation, called MEME (Bailey and Elkan 1995).

Gibbs sampling is the stochastic variant of the EM algorithm (Lawrence et al. 1993). Because the stochastic nature of the Gibbs sampling methods, they do not suffer from getting stuck in a local optimum. On the other hand, due to this stochastic nature, the algorithm might need many iterations to obtain adequate results. Currently, there exist different implementations of Gibbs sampling, each with different extra features such as automatic detection of motif length, use of a higher-order background model to describe DNA sequences, etc: AlignACE (Roth et al. 1998), ANN-Spec (Workman and Stormo 2000), BioProspector (Liu et al. 2001) and MotifSampler (Thijs et al. 2001; Thijs et al. 2002a; Thijs et al. 2002b).

1.4.2.1.2 Combinatorial methods

This second category of motif detection procedures uses a string-based representation of a regulatory motif. Most commonly are algorithms that enumerate words up to a maximum size; these words (i.e., motifs) are then scored with an appropriate measure of statistical significance. Examples of such word-counting methods have been described by Brazma et al. (1998), Van Helden et al. (1998) and Pavesi et al. (2004b; 2006)

1.4.2.2 Phylogenetic footprinting

Phylogenetic footprinting (PF) is a comparative methodology that uses cross-species sequence conservation to identify regulatory motifs (Tagle et al. 1988; Gumucio et al. 1992; Duret and Bucher 1997; Zhang and Gerstein 2003; Ureta-Vidal et al. 2003). PF is based on the observation that functional elements evolve more slowly under selective pressure than the non-functional background sequence. As a consequence functional elements are conserved among evolutionary divergent genomes. Based on this observation PF detects conserved regions in orthologous non-coding DNA sequences (i.e., intergenic sequences of genes that are the result of speciation); these regions are likely to be/contain regulatory motifs (Figure 1-7). PF has proven successful for the identification of regulatory motifs in prokaryotes (McGuire et al. 2000; McCue et al. 2001; Marchal et al. 2004) and yeast (Cliften et al. 2003; Kellis et al. 2003; Kellis et al. 2004). Furthermore, several vertebrate studies have been published (Wasserman and Fickett 1998; Jareborg et al. 1999; Wasserman et al. 2000; Loots et al. 2000; Krivan and Wasserman 2001; Gottgens et al. 2002; Muller et al. 2002; Santini et al. 2003; Lemos et al. 2004; Boutros et al. 2004; Givens et al. 2004).

In order to identify such evolutionary conserved regions, orthologous sequences are aligned (see §1.4.2.2.1). To perform these alignments there exist two types of methods: the global and the local

alignment methods (Ureta-Vidal et al. 2003). These are discussed in §1.4.2.2.2 respectively §1.4.2.2.3.

Besides methodologies developed to identify evolutionary conserved regions, there also exist databases that contain currently identified conserved regions. A few examples are given in §1.4.2.2.4.

1.4.2.2.1 Introduction to sequence alignment

When aligning sequences, sequences are written beneath each other; identical or similar characters (nucleotides or amino acids) are placed in the same column; characters that are not similar can be placed in the same column as 'mismatch' or against a gap in the other sequence. In an optimal alignment non-similar characters and gaps are placed in such manner that the maximal number of identical or similar characters are placed in the same column.

The total score of an alignment is calculated as the sum of similarity scores for each aligned pair of characters. These similarity measures are represented in a score- or substitution matrix, such as for instance BLOSUM (for protein alignment) (Henikoff and Henikoff 1994). The occurrence of gaps in the alignment is often penalized with a negative gap penalty.

Besides scoring the alignment, there is need for an algorithm that finds the optimal alignment. In the next two paragraphs an overview is given of the most important algorithms for generating pairwise and multiple alignments, i.e., comparing two respectively multiple sequences.

We like to draw the attention that both protein and nucleotide sequences can be the subject of alignment. However, in the scope of this research -the quest for regulatory motifs- only alignment of DNA sequence is relevant and therefore only these types of alignment algorithms will be discussed.

1.4.2.2.2 Global alignment methods

Pairwise global alignment methods

One of the oldest and most popular pairwise alignment methods is the Needleman-Wunsch (NW) algorithm (Needleman and Wunsch 1970). This algorithm is based on dynamic programming (DyP), i.e., the solution is reached by combining the solutions to subproblems. It can be mathematically proven that a DyP-algorithm always finds the optimal alignment and this is thus also the case for NW. However, the scoring scheme used in the NW-algorithm was primarily developed to compare protein sequences. Because of its computational complexity, it is not well suited to align long stretches of DNA (e.g., contigs).

Therefore, new heuristic global alignment algorithms were developed in the light of the genome projects (see §1.1). These algorithms have a higher performance for detecting conserved regions in genomic sequences, i.e., long DNA sequences characterized by alterations of conserved exons and more variable introns. All these algorithms assume that aligned sequences are completely colinear, meaning that no inversions or translocations have taken place (Brudno and Morgenstern 2002; Ureta-Vidal et al. 2003). As mentioned these algorithms are heuristic and therefore they offer no guaranty of finding the optimal alignment. On the other hand, their high performance enables them to align complete genome sequences. Although these heuristic methods were primarily developed to compare coding regions, they were already successfully applied for detection of conserved non-coding (intergenic) regions. In what follows we give an overview of the most widely used global alignment methods that were proven successful in aligning intergenic sequences. Given the multitude of alignment algorithms developed and published, it is inevitable to miss a few alignment algorithms; our objectives were merely to provide an overview of the main contributions in this area.

All the heuristic methods mentioned, use the same working scheme. They all involve three steps, illustrated in Figure 1-6. Step 1, the seeding step, involves identification of all conserved subsequences, also referred to as matches. These matches are then used to anchor the sequences onto each other in step 2. In the final step, the sequences regions between the anchors are aligned suing a DyP-algorithm (for instance NW) (Ureta-Vidal et al. 2003). For most implementations, steps 1 through 3 are recursively repeated. Furthermore distinct algorithms differ in seeding strategy.

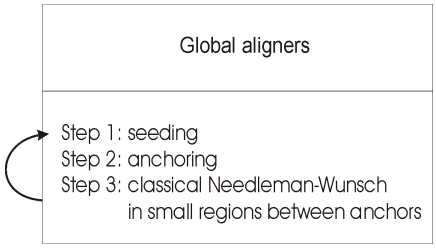


Figure 1-6. Schematic representation of the genome wide global alignment strategy. Step 1, the seeding step, involves identification of all conserved subsequences, i.e., matches. In step 2 the sequences are anchored onto each other by means of the matches identified in step 1. Finally, in step 3, the regions between the anchors are aligned using a dynamic programming (DyP) algorithm. As the curved arrow indicates, (in most algorithms) these three steps are recursively repeated. Adapted from: Ureta-Vidal et al. (2003).

A widely used alignment algorithm is AVID, which uses suffix trees in its seeding step (Bray et al. 2003).

LAGAN uses a strategy analogous to AVID, but seeding is performed using CHAOS (Brudno and Morgenstern 2002; Brudno et al. 2003a; Brudno et al. 2003b). In contrast to suffix trees (AVID), CHAOS searches for short inexact matches. Therefore, LAGAN is better suited to align strongly diverged species. Furthermore, this algorithm has been shown to perform better in aligning long intergenic sequences (Ureta-Vidal et al. 2003; Brudno et al. 2004).

Multiple global alignment methods

Initially new pairwise alignment methods were developed to compare for instance, human with mouse, but with the increasing availability of vertebrate genomes (see §1.1) a shift towards multiple alignment methods was indispensable (Brudno et al. 2004). Furthermore, the incorporation of more divergent genomes improves the detection of regulatory motifs (Boffelli et al. 2003): by augmenting the phylogenetic diversity of the sequences studied it becomes easier to distinguish between conserved regulatory motifs and variable, non-functional background sequence. This is illustrated in Figure 1-7.

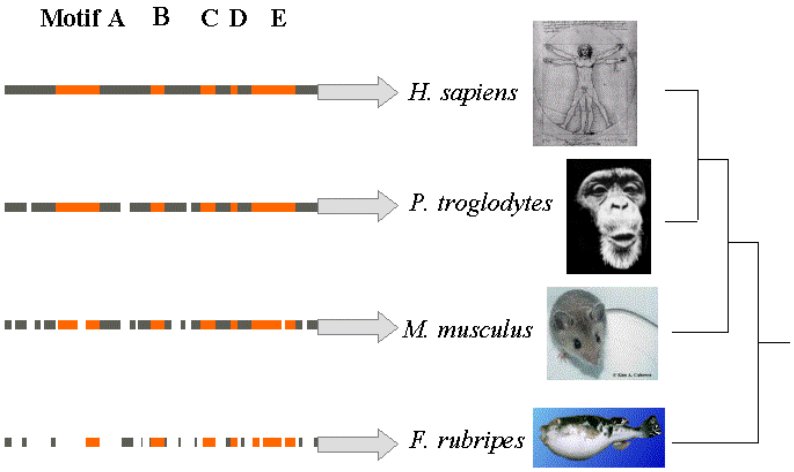


Figure 1-7. Schematic representation of phylogenetic footprinting. PF detects regulatory motifs by identification of conserved regions in the intergenic regions of orthologous genes. As is depicted in the illustrations, the more species are included in the analysis and the further evolved they are, the easier it becomes to differentiate between (conserved) regulatory motifs and not-conserved non-functional background sequence. Indeed, more evolutionary time has elapsed, giving the time to accumulate mutations. Symbol code: motifs: orange, non-functional intergenic sequence: dark grey; coding genes: light grey arrows.

Heuristic multiple global alignments are assembled from the pairwise alignments using a guide tree. This principle of progressive alignments is used, for example, in CLUSTALW (Thompson et al. 1994; Wasserman and Fickett 1998) and MAVID (Bray and Pachter 2003; Bray and Pachter 2004).

LAGAN was also extended to a multiple alignment implementation, Multi-LAGAN or MLAGAN, that also uses progressive alignment (Brudno et al. 2003b). However, this algorithm offers an optional iterative step: consecutively, each sequence is removed from the multiple alignment and then re-aligned with the remainder of the alignment. This step is repeated until no further improvement is observed (Brudno et al. 2003b). This iterative step is an example of the principle of iterative refinement.

DIALIGN is an heuristic alignment algorithm that uses an alternative approach to progressive alignment (Morgenstern et al. 1998; Morgenstern 1999): pairwise conserved regions, 'blocks', are assembled to a multiple alignment by the use of a 'greedy'-algorithm (Pollard et al. 2004). Even though DIALIGN has proven successful for detection of regulatory elements (PF) in vertebrate sequences (Gottgens et al. 2001; Gottgens et al. 2002), it is mostly qualified to align protein sequences and relatively short DNA sequences. Brudno et al. (2003a) extended the application of DIALIGN by combining it with the formerly mentioned CHAOS. The fast algorithm CHAOS is used to identify subsequences with high similarity (Brudno and Morgenstern 2002). These matches are then used as anchors to generate the final global alignment (as explained for AVID and LAGAN).

Recently, Siddharthan (2006) developed an algorithm Sigma, which is similar to DIALIGN but which uses a scoring scheme specifically for non-coding DNA. Indeed, the algorithm uses a background model derived from actual DNA, for instance, intergenic sequences.

1.4.2.2.3 Local alignments methods

Besides the global aligners, local alignment algorithms are also used to detect phylogenetically conserved regions. These implementations calculate an optimal similarity between subregions of the aligned sequences.

Smith and Waterman, for instance, developed a local version of the NW algorithm that enables alignment of sequences that are strongly similar over only a limited distance (Smith and Waterman 1981). But the resulting alignment is strongly dependent on the choice of parameters: if a high penalty is assigned to gaps, the alignment will be focused on only one or two strong conserved regions. A low gap penalty, on the other hand, will result in detection of multiple, less conserved regions. Another drawback of the Smith-Waterman (SW) algorithm is its computational intensiveness, which

makes it inappropriate for aligning of long sequences (Ureta-Vidal et al. 2003).

BLASTZ can be used for the local alignment of long genomic sequences (Schwartz et al. 2000; Schwartz et al. 2003). This algorithm consists of three steps, analog to the higher mentioned heuristic global alignment algorithms (Gilligan et al. 2002) .

Analogous to global alignment algorithms (see §1.4.2.2.2), both SW and BLASTZ assume that the aligned sequences are colinear. This demand is too stringent for detection of regulatory motifs: although the sequences of regulatory motifs are often conserved in the intergenic sequences of orthologs (principle of PF), this does not imply that the order in which the regulatory motifs occur is also maintained; i.e., the assumption of colinearity does not necessarily hold.

Therefore multiple local alignment algorithms that do not assume conservation of order are better suited for identification of regulatory motifs and modules. Motif detection algorithms, as discussed in §1.4.2.1, are candidate methodologies. However, these motif detection algorithms were initially developed to study independent sequences, namely co-regulated genes that are not phylogenetically related to each other. This assumption of independence is in contradiction with the fundaments of PF: because of the high degree of sequence conservation between the intergenic regions of orthologs of closely related species, the number of local optima (i.e., conserved subsequences) will be high. This makes it more difficult to distinguish between biological relevant conserved regulatory motifs and the artifacts (false positives) due to overall sequence conservation (Marchal et al. 2004).

Another category of algorithms identifies evolutionary conserved motifs in a set of orthologous sequences, taking into account the phylogenetic relationships between the sequence, e.g., FootPrinter (Blanchette et al. 2000; Blanchette and Tompa 2003). This string-based implementation (see §1.4.2.1.2) identifies DNA motifs that have evolved more than the background sequence (Blanchette and Tompa 2002). Using DyP, FootPrinter searches the phylogenetic tree from leafs to root, searching for motifs that show a minimal number of mismatches (Rombauts et al. 2003). A drawback of this combinatorial method is that is not suited to analyze long sequences.

1.4.2.2.4 Precomputed alignments

Several groups have performed large cross-species comparisons and made them publicly available through genome browsers: here, researchers can download or visualize previously identified evolutionary conserved regions (Ureta-Vidal et al. 2003). A few commonly used databases

containing vertebrate genomes are ECR (evolutionary conserved regions) browser, Ensembl (Birney et al. 2006), Vista genome browser (Mayor et al. 2000; Frazer et al. 2004), UCR (ultra-conserved regions) browser and UCSC genome browser (Kent et al. 2002; Karolchik et al. 2003).

1.4.2.3 *De novo* detection of regulatory modules

As was explained in §1.4.1.2 most approaches to predict regulatory modules can be classified as motif screening methods, i.e., dependent on the availability of known motifs. In the framework of *de novo* identification tools, we would like to mention the existence of a limited number of *de novo* motif module discovery algorithms, such as CisModule (Zhou and Wong 2004) and EMCMODULE (Gupta and Liu 2005). The latter method identifies modules using a collection of motif models that are obtained from both *de novo* motif search using existing algorithms and databases of known TFBSs.

1.5 Outline thesis

1.5.1 Problem statement

In order to understand how cells (and organisms) behave in specific circumstances it is necessary to reveal the interactions that take place within the cells, such as protein-protein interactions, binding of proteins to DNA, DNA-DNA interactions, etc…. One part of the puzzle lies in the gene regulatory networks (GRNs): as explained in §1.3.3 GRNs enclose the interplay between regulatory motifs (also organized in modules) and TFs and its impact on gene expression.

Since regulatory motifs are important building stones of these GRNs, the genome wide localization and functional assessment of TFBSs would be a major leap towards unraveling such networks. As mentioned before, our current knowledge of regulatory motifs is miniscule given the large amount of regulatory motifs expected to function in, for instance, vertebrate genomes.

As we demonstrated in the previous paragraph (§1.4) there already exist many possible strategies to identify regulatory motifs: a first approach is searching for known regulatory motifs, this is referred to as motif screening (§1.4.1). Alternatively, their exist methodologies that search for novel not-yet characterized regulatory motifs. Such *de novo* motif identification methods (§1.4.2) can be divided in two categories that differ in the type of sequence data they use: motif detection algorithms (§1.4.2.1)

start from co-expressed genes assuming that a common TF-motif interaction is responsible for the common expression pattern. The second type of *de novo* motif identification is phylogenetic footprinting (PF, §1.4.2.2), which starts from orthologous sequences: non-coding DNA motifs conserved over evolution are expected to be functional and are thus likely to be/contain regulatory motifs.

With al these different motif discovery approaches it almost seems as if the identification of regulatory motifs is merely a question of time; time to apply the appropriate algorithm to the appropriate dataset. This, however, is not the case, especially for motif discovery in vertebrate organisms (Tompa et al. 2005): as is discussed in the remainder of this paragraph, each approach has its own advantages and drawbacks. Since this thesis focuses on motif detection in vertebrate organisms, we here specifically emphasize the limitations of each approach for motif detection in vertebrates; these are also summarized in Table 1-2.

Table 1-2. The main disadvantages of the different types of motif discovery approaches for motif identification in vertebrate organisms. For each type of approach, the paragraph is indicated were this type of methodologies is explained in more detail (§) followed by the key limitations of this approach for motif detection in vertebrate sequences (limitations).

Approach	§	Limitations
Motif screening	1.4.1	limited number of known regulatory motifs many false positives (certainly in long vertebrate intergenic regions)
Motif detection (local alignment)	1.4.2.1	many false positives in long vertebrate intergenic regions
Phylogenetic footprinting (global alignment)	1.4.2.2	heterogeneity in length of vertebrate intergenic regions hinders alignment using global alignment strategies

First, the applicability of motif screening procedures is restricted due to the currently limited knowledge of TFBSs (Table 1-2). Motif screening algorithms are useful, if one is interested, for instance, in one specific regulator and wishes to recover all potential binding sites for this regulator. Nevertheless, this type of algorithms will never lead us to revealing novel, not yet characterized TFBSs. Furthermore, motif screening results in

detection of many false positives, i.e., motifs that do not have any biological function. The fact that vertebrate genes are typically characterized by long intergenic regions, ranging up to several hundreds kb, makes motif screening procedures inappropriate for motif recovery in vertebrate intergenic sequences.

In order to expand our knowledge of regulatory motifs, we are thus forced to use *de novo* identification tools (§1.4.2).

The application of motif detection procedures (i.e., local alignment) in higher vertebrates is also limited because of the long vertebrate intergenic regions. Given the small nature of regulatory motifs (5 to 8 bp) this results in a low signal-to-noise ratio, leading to many false positive motif hits.

As explained before, inclusion of multiple, more distantly related species can greatly improve the detection of conserved regulatory motifs by PF (Figure 1-7). However, vertebrate intergenic regions may differ considerably in size, for instance, between mammals and the pufferfish *Fugu* (Venkatesh et al. 2000; Aparicio et al. 2002). This heterogeneity in sequence size together with a low overall conservation (due to long evolutionary distance) hinder the correct alignment using global alignment algorithms (§1.4.2.2.2).

In conclusion, neither of the existing strategies for motif discovery is optimally suited to identify novel motifs in distantly related vertebrate organisms.

1.5.2 Overview of the thesis

Because neither of the existing motif recovery strategies is ideal for motif identification in strongly diverged vertebrate organisms, in this thesis we combined the different approaches. This is schematically represented in Figure 1-8. By incorporating different motif discovery approaches in our research, we take advantage of each of the methods used. Furthermore, by applying each strategy on a well-suited data type we minimize its drawbacks (summarized in Table 1-2). For instance, applying motif detection tools (i.e., local alignment algorithms) on shorter sequences will reduce the number of false positive motifs compared to applying it to full-length intergenic sequences.

This logic forms the basis of the novel motif detection strategy we developed in **chapter 2**: we developed a new procedure that combines the advantages of two different existing strategies: phylogenetic footprinting (§1.4.2.2) and motif detection (§1.4.2.1). This methodology consists of two steps, which explains why we refer to it as ‘two-step procedure’ (see Figure 1-9), and was developed to identify regulatory motifs in distantly related

species. In the first step, we use a global alignment algorithm to identify regions that are conserved among closely related species. In this way we reduce the amount of sequence data and thus increase the signal-to-noise ratio (motif vs. background sequence) for the next step. Because the stretches of DNA searched are shorter (compared to the initial intergenic regions) we are now able to use a motif detection algorithm to identify regulatory motifs. For the second, motif detection step we developed a new implementation, BlockSampler, which is based on Gibbs Sampling. This implementation searches for long conserved blocks (cfr. motifs) instead of looking for short motifs. In this way we minimize the number of false positive hits. In conclusion, by combining global alignment (step 1: data reduction) and local alignment (step 2: motif detection step) we developed a strategy that enables phylogenetic footprinting in distantly related species (Figure 1-8, Figure 1-9). Table 1-3 summarizes how we made the existing strategies (motif detection and phylogenetic footprinting) applicable to distantly related species.

The resulting two-step approach was then tested on four benchmark datasets and we compared its performance with that of alternative approaches (motif detection and global alignment methods). This showed that our two-step approach clearly outperforms alternative methodologies when the intergenic sequences become either too long (typical for vertebrate genes) or too heterogeneous in size (cfr. strongly evolved species). Furthermore, we showed that the strength of the methodology lies in the combination of global and local alignment, i.e. data reduction and motif detection.

Application of our methodology to additional datasets showed that the two-step approach is applicable to a variable number of sequences related by variable evolutionary distances, thus providing a generally applicable methodology.

The development and testing of the two-step approach was published in Genome Biology in 2005:

Van Hellemont R, Monsieurs P, Thijs G, De Moor B, Van de Peer Y and Marchal K. A novel approach to identifying regulatory motifs in distantly related genomes. Genome Biology, 2005, 6 (13): R113.1-R113.18.

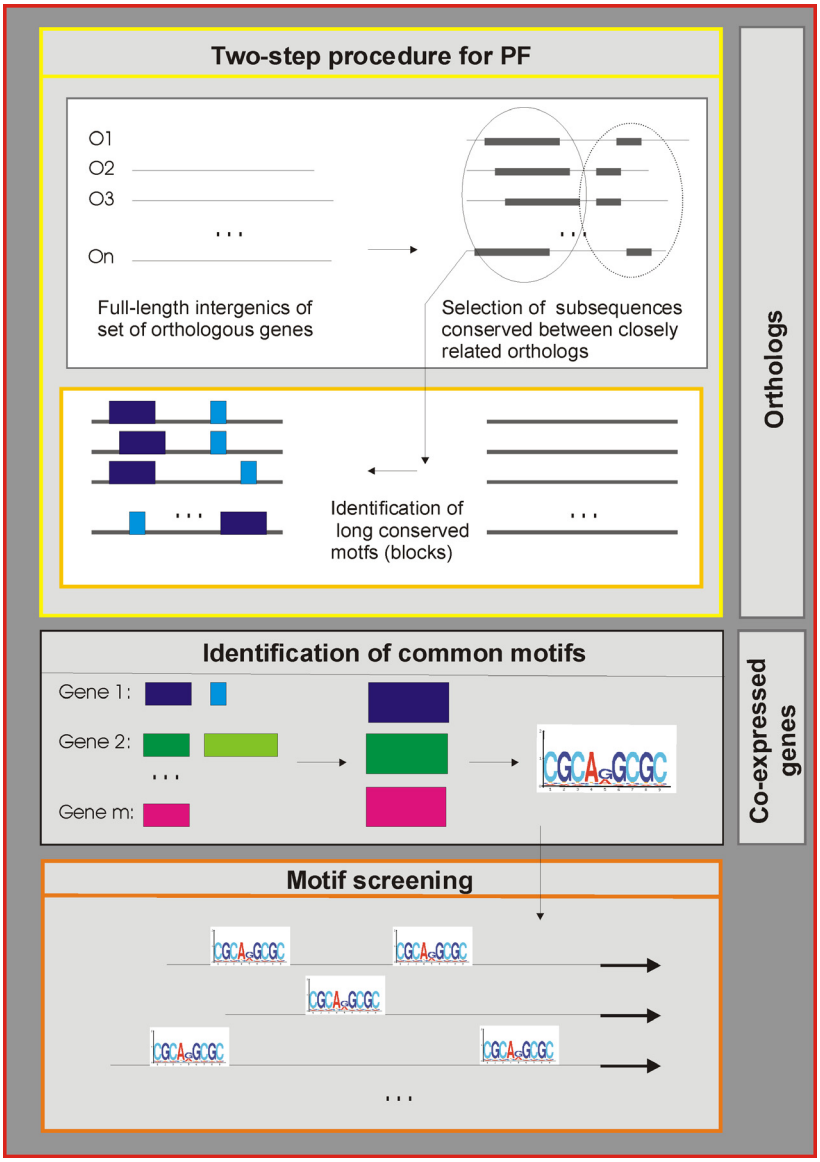


Figure 1-8. Integration of different motif discovery strategies. Because none of the existing motif recovery strategies is ideally suited for motif identification in vertebrate organisms, in this thesis we combined the different approaches. The two-step approach (indicated with yellow square) enabled motif detection in orthologous vertebrate intergenic sequences (phylogenetic footprinting, PF) by combining global alignment with local alignment (motif detection). Identification of reliable motifs in co-expressed genes, which is normally performed with motif detection algorithm, was successfully done by combining the developed (two-step) and a clustering step. Finally, motif screening was used to reveal additional motif instances of the identified motifs. This scheme is also used in Figure 1-9 that gives a structural overview of the thesis based on the methodology used.

Chapter 3 describes a first biological application of our newly developed methodology. The goal of this study was to identify regulatory motifs that are in support of subfunctionalization. This research was performed in collaboration with research division 'Bioinformatics and Evolutionary Genomics' (Department of Plant Systems Biology, University of Ghent) under supervision of Prof. Yves Van de Peer.

Subfunctionalization indicates that following a duplication event the resulting duplicates (paralogs) divide the gene's original functions. Since the presence of both paralogs is necessary to guarantee a flawless functioning of the organism both gene copies will be retained in the organisms genome. Since the activity of a gene is largely determined by its transcriptional regulation, differences in regulatory motifs between the two paralogs can, for instance, be responsible for distinct expression patterns that together sum up to the expression pattern of the original gene (i.e., subfunctionalization). This theory is experimentally supported by a (albeit limited) number of studies that have revealed regulatory motifs that are divergently retained between paralogs in accordance with the divergent expression patterns between these paralogs. However, there exists no methodology to identify regulatory motifs that are in support of subfunctionalization. Our challenge was to develop such a methodology that is not only generic but also applicable on a genome-wide scale.

In order to identify regulatory motifs responsible for such a divergent expression pattern, we applied the newly developed BlockSampler to sets of orthologous sequences containing fish paralogs (Figure 1-9). Doing so, we were able to identify regulatory motifs that are divergently conserved among the duplicates. By including sets of paralogs known to exhibit a divergent expression pattern, we were able to interpret the results. In the majority of these proof-of-concept datasets we were able to identify motifs that support the experimentally observed expression divergence.

The results of this study show that BlockSampler is very well suited to identify reliable motifs that support subfunctionalization. Moreover, it is possible to use BlockSampler on a genome wide scale.

This results discussed in chapter 3 have been published as book chapter in Genome Dynamics vol. III on 'Gene and Protein Evolution':

Van Hellemont R, Blomme T, Van de Peer Y and Marchal K. Divergence of regulatory sequences in duplicated fish genes. Invited book chapter in 'Genome Dynamics vol. 3: Gene and Protein Evolution**',** 2007 (in press). Editor: Volff, J.-N.

Table 1-3. The solutions provided in this thesis that enable the use of the different motif discovery approaches for motif identification in strongly diverged vertebrate organisms. For each type of approach, the key limitations of this approach for motif detection in vertebrate sequences are indicated (limitations) followed by the solution provided in this thesis that make the strategies applicable to vertebrate organisms.

Approach	Limitations	Solutions provided in this thesis
Motif screening	limited number of known regulatory motifs many false positives (certainly in long vertebrate intergenic regions)	*de novo* motif detection + motif screening: => identification of novel TFBSs
Motif detection (local alignment)	many false positives in long vertebrate intergenic regions	Data reduction: preselection of regions conserved among closely related species => increase signal-to-noise ratio => less false positives Motif detection: BlockSampler identifies long conserved blocks instead of short motifs => less false positives
Phylogenetic Footprinting	heterogeneity in length of vertebrate intergenic regions hinders alignment using global alignment strategies	Data reduction: preselection of conserved intergenic for species characterized by long intergenic regions diminishes the differences in length between vertebrates (e.g., mammals and pufferfish) Motif detection: finding motifs with the optimal motif length (long blocks)

As a consequence of the existing productive collaboration between SISTA-SCD and Legendo (group of Prof. Roger Bouillon), an important part of this thesis is dedicated to the study of the mechanism of action of vitamin D_3. Besides its classical actions in mineral homeostasis, 1,25-dihydroxyvitamin D_3 (vitD3), the active metabolite of vitamin D_3, has been shown to exert some growth-inhibitory effects. In order to gain further insight in the molecular mechanism behind both the classical and the non-classical effects, we performed two different types of analyses that are the subject of chapters 4 and 5. In both studies we started from sets of genes showing a similar expression pattern after treatment with vitD3; we assume that a common regulatory motif is at the basis of their co-expression.

In **chapter 4**, we screen the promoter regions of the co-expressed genes with known TFBSs from which we expect that they could be responsible for the observed expression pattern (Figure 1-8, Figure 1-9). This screening led to identification of E2F as an important player in the vitD3-induced growth-inhibition. Moreover, we were able to identify a number of previous unknown E2F target genes.

The results of this study were published in The Journal of Biological Chemistry in 2005

Verlinden L, Eelen G, Beullens I, Van Camp M, Van Hummelen P, Engelen K, Van Hellemont R, Marchal K, De Moor B, Foijer F, Te Riele H, Beullens M, Bollen M, Mathieu C, Bouillon R and Verstuyf A. Characterization of the condensin component Cnap1 and protein kinase Melk as novel E2F target genes down-regulated by 1,25-dihydroxyvitamin D_3. 2005. The Journal of Biological Chemistry, 280 (45): 37319-37330.

and in The Journal of Steroid Biochemistry and Molecular Biology in 2006

Verlinden L, Eelen G, Van Hellemont R, Engelen K, Beullens I, Van Camp M, Marchal K, Mathieu C, Bouillon R and Verstuyf A. 1alpha,25-Dihydroxyvitamin D(3)-induced down-regulation of the checkpoint proteins, Chk1 and Claspin, is mediated by the pocket proteins p107 and p130. The Journal of Steroid Biochemistry and Molecular Biology, 2006, in press.

In **chapter 5**, we study the presence of novel, not yet characterized regulatory motifs responsible for the co-expression of vitD3-regulated genes. Indeed, because only a fraction of the TFBSs in vertebrates is known, it is likely that the similar expression pattern of the co-expressed genes (clusters of vitD3-regulated genes) is due to the binding of a TF to a yet unidentified regulatory motif (cfr. *de novo* motif discovery algorithms §1.4.2). In order to do this we developed a new motif discovery strategy that combines the two-step approach (Figure 1-8 and Figure 1-9, Two-step procedure for PF) and the methodology developed by Monsieurs et al. (2006). The latter method exploits both orthology and co-expression information in order to discover *de novo* motifs. This methodology, however, was developed for the identification of motifs in prokaryotic sequences. In order to apply this strategy to vertebrate sequences, we used the two-step procedure for the phylogenetic footprinting step.

As explained above, the two-step approach enables us to identify conserved motifs in orthologous sequences. Using this approach we identified evolutionary conserved motifs for each individual gene. Next, we assessed whether we could find motifs that are evolutionary conserved in multiple co-expressed genes (i.e., sets of orthologs for that gene). This last step is indicated in Figure 1-8 as 'Identification of common motifs'.

We applied this novel methodology to a set of genes that is quickly up-regulated after cells have been treated with vitD3. This revealed 31 motifs that are possibly (at least partially) responsible for the up-regulated expression pattern. Some of them correspond to known TFBSs, enabling us to interpret the results. Because many assumptions have been made in the three steps leading to the discovery of these 31 motifs, it is likely that we missed a few motifs. Because, it is important to have a more or less accurate estimate of the number of (reliable) motifs in order to measure its input to the common expression pattern, we finally screened the intergenic sequences of all the co-expressed genes with the each of the 31 motif models (Figure 1-8, Figure 1-9: Motif screening). This revealed many more motif hits. Further experimental work is necessary to validate and interpret these results.

By combing *de novo* motif detection (two-step procedure) with motif screening, we were able to use motif screening for the identification of novel, yet-unknown TFBSs (Table 1-3).

A schematic representation of the structure of this thesis is illustrated in Figure 1-9: this diagram depicts the methodologies developed and links these with the distinct chapters. The main achievements of this work are also summarized in Table 1-4.

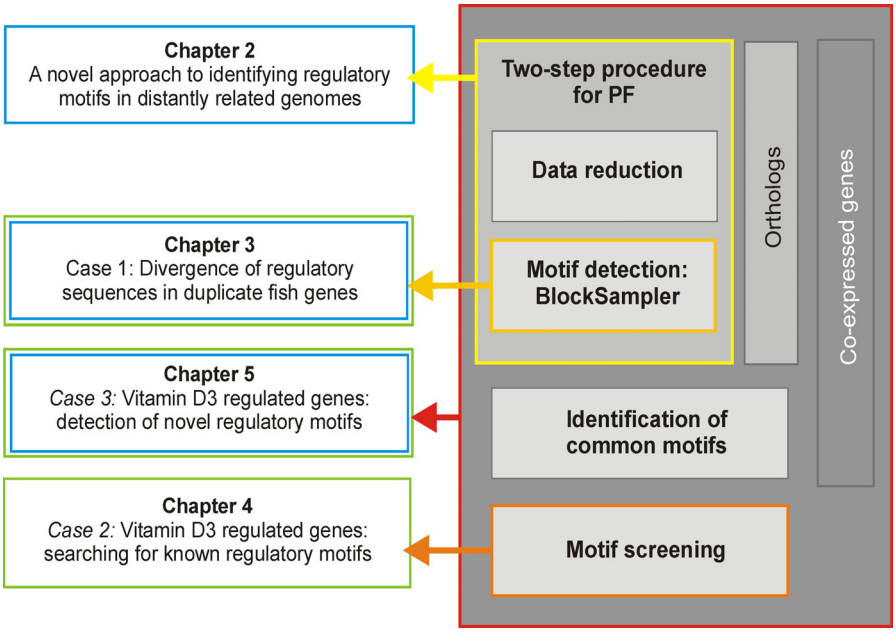


Figure 1-9. Overview of the thesis. Chapter 2 explains the development and application of a novel two-step procure for phylogenetic footprinting. Chapter 3 describes the application of BlockSampler for the detection of regulatory motifs in support of subfunctionalization. The subject of chapter 4 is the application of motif screening methods to identify known regulatory motifs in sets of vitamin D_3-regulated genes. Chapter 5 describes the application of a novel methodology to identify evolutionary conserved regulatory motifs in sets of co-expressed genes and the results of applying this methodology to a set of vitamin D_3-regulated genes. Color code: blue: methodological achievements (also see Table 1-4); green: biological achievements (also see Table 1-4); yellow to red: methodologies applied to reach achievements indicated in Table 1-4; grey shades: overview and incorporation of the methods developed and/or applied.

Table 1-4. Achievements in this work. Chapter: the chapter reference. Type: the type of the achievement: developing a novel methodology (M; indicated in blue in Figure 1-9**)** or solving biological problems (B; indicated in green in Figure 1-9). Achievement: a short description of the achievement. Pub: the corresponding article where the study is published.

Chapter	Type	Achievement	Pub
2	M	Generic two-step procedure for phylogenetic footprinting in distantly related vertebrate organisms	Van Hellemont et al. A novel approach to identifying regulatory motifs in distantly related genomes. 2005. Genome Biology, 2005, 6 (13): R113.1-R113.18.
3	M	First generic methodology to identify regulatory motifs in support of subfunctionalization on a genome wide scale	Van Hellemont R, Blomme T, Van de Peer Y and Marchal K. Divergence of regulatory sequences in duplicated fish genes. Invited book chapter in 'Genome Dynamics vol. 3: Gene and Protein Evolution', 2007 (in press). Editor: Volff, J.-N..
	B	Identification of motifs that support experimentally observed expression divergence	
4	B	Detection of known transcription factor binding sites in sets of co-regulated genes: identification of E2F as an important player in the vitD3-induced growth-inhibition.	Verlinden et al. Characterization of the condensin component Cnap1 and protein kinase Melk as novel E2F target genes down-regulated by 1,25-dihydroxyvitamin D_3. 2005. The Journal of Biological Chemistry, 280 (45): 37319-37330. Verlinden L, Eelen G, Van Hellemont R, Engelen K, Beullens I, Van Camp M, Marchal K, Mathieu C, Bouillon R and Verstuyf A. 1alpha,25-Dihydroxyvitamin D(3)-induced down-regulation of the checkpoint proteins, Chk1 and Claspin, is mediated by the pocket proteins p107 and p130. The Journal of Steroid Biochemistry and Molecular Biology, 2006, in press.
5	M	Generic methodology to identify evolutionary conserved motifs shared by multiple co-regulated genes and thus possibly responsible for their similar expression pattern	
	B	Identification of 31 conserved motifs likely to be responsible for vitD3-induced expression pattern	

Chapter 2

A novel approach to identify regulatory motifs in distantly related genomes

2.1 Introduction

In this chapter we will develop a methodology to identify evolutionary conserved motifs in eukaryotic sequences. This chapter has been published in Genome Biology (Van Hellemont et al. 2005). Additional information is given in Appendix A; where relevant this is indicated in the text.

Phylogenetic footprinting is a comparative method that uses cross-species sequence conservation to identify new regulatory motifs (Tagle et al. 1988). Based on the observation that functional regulatory motifs evolve more slowly than non-functional sequences, the method identifies potential regulatory motifs by detecting conserved regions in orthologous intergenic sequences (Fickett and Wasserman 2000; Levy et al. 2001). The comparison of orthologous sequences from multiple genomes is often based on multiple sequence alignment, (Boffelli et al. 2003; Chapman et al. 2004) and several alignment algorithms, such as CLUSTALW (Thompson et al. 1994), DIALIGN (Morgenstern et al. 1998; Morgenstern 1999), MAVID (Bray and Pachter 2003; Bray and Pachter 2004) and MLAGAN (Brudno et al. 2003) have proven very useful to identify conserved motifs in closely related higher vertebrate sequences (Wasserman and Fickett 1998; Boffelli et al. 2003; Major and Jones 2004). Although the comparison of closely related organisms has proven successful, inclusion of more distantly related species can greatly improve the detection of conserved regulatory motifs. By adding more distantly related sequences, the conserved functional motifs can be more easily distinguished from the often highly variable 'background' sequence. Moreover, this leads to the detection of motifs that have a function in a wider variety of organisms, e.g., all vertebrates (Aparicio et al. 1995; Bagheri-Fam et al. 2001; Montpetit and Sinnett 2001; Abrahams et al. 2002;

Santini et al. 2003; Nobrega et al. 2003). Both Sandelin et al. (2004) as Woolfe et al. (2005), for instance, performed a whole genome comparison of human and pufferfish, which diverged approximately 450 mya to discover non-coding elements conserved in both organisms. They showed that most of these conserved non-coding elements are located in regions of low gene density (implying long intergenic regions) (Woolfe et al. 2005). Moreover, many of the conserved non-coding elements are located at large distances from the nearest gene (Sandelin et al. 2004; Woolfe et al. 2005). These findings lead to the conclusion that it is interesting to analyze whole intergenic regions of vertebrate genes, rather than limit the comparative analyses to the promoter region located near the transcription start.

However, vertebrate intergenic regions may differ considerably in size, for instance, when comparing intergenics of e.g. mammals with those of *Fugu* (Brenner et al. 1993; Venkatesh et al. 2000; Aparicio et al. 2002). Since multiple sequence alignments are often based on global alignment procedures, they will likely fail to correctly align such sequences of heterogeneous length (Elemento and Tavazoie 2005).

An alternative for alignment methods is the use of *de novo* motif detection procedures for phylogenetic footprinting. These are based on either probabilistic or combinatorial algorithms. One such method, i.e., FootPrinter, uses a string based motif representation with dynamic programming to search a phylogenetic tree for motifs that show a minimal number of mismatches (Blanchette and Tompa 2002; Blanchette and Tompa 2003). Probabilistic algorithms, such as MEME (Bailey and Elkan 1995), Consensus (Hertz et al. 1990; Hertz and Stormo 1999), and Gibbs sampling (Lawrence et al. 1993; McCue et al. 2001), use a matrix representation of the motif (position specific weight matrix). Currently, several implementations of Gibbs sampling are available, such as AlignACE (Hughes et al. 2000; Cliften et al. 2001), ANN-spec (Workman and Stormo 2000), BioProspector (Liu et al. 2001) and MotifSampler (Thijs et al. 2001; Thijs et al. 2002a; Thijs et al. 2002b; Tompa et al. 2005). However, these algorithms are sensitive to low signal-to-noise ratios, i.e. the presence of small motifs (5-8 bp) in long intergenic sequences. This often results in the detection of many false positive motifs. On the other hand, an advantage of these procedures is that, because motif detection comes down to locally aligning the orthologous sequences, non-collinear motifs can still be detected.

In conclusion, neither motif detection nor multiple alignment methods are optimally suited to correctly align long intergenic sequences of heterogeneous length. Here, we present a simple two-step procedure that identifies conserved regions by combining the advantages of both alignment and motif detection methods. Such highly conserved regions most likely contain transcription factor binding sites or other functional intergenic sequences (Pennacchio 2003). To show its efficiency, we applied our two-

step approach to well described benchmark datasets. Since regions of strong conservation among divergent vertebrates are often associated with developmental regulators (Sandelin et al. 2004; Woolfe et al. 2005) we choose mainly this type of genes to test our methodology. The presented approach, however, is applicable to any set of organisms and genes for which one wants to compare the intergenic sequences.

2.2 Results

2.2.1 A two-step procedure for phylogenetic footprinting

In this study we aim to detect regulatory motifs that have been retained over long periods in evolution; in our test case this applied to mammals to ray-finned fishes such as *Fugu*. The *Fugu* genome, however, is very compact and approximately 8 or 9 times smaller than the human one, although both genomes are assumed to contain a similar repertoire of genes. The compactness of the genome of *Fugu* is a result of shorter intergenic regions and introns (Brenner et al. 1993; Elgar et al. 1996; Aparicio et al. 2002). On the other hand, the preliminary and still often erroneous annotation of the *Fugu* genome sometimes results in the selection of very long intergenic regions. Such heterogeneous size of the intergenic regions that need to be compared complicates identification of regulatory motifs. Widely used alignment algorithms, such as AVID, LAGAN, and others, will usually fail when the sequences that need to be aligned differ too drastically in length. This problem is exacerbated when the sequences have a low overall percent identity. To cope with this, motif detection procedures could offer a solution. However, because regulatory motifs are typically only 6 to 30 bp long, whereas intergenic sequences of vertebrate genes range up to tens of kilobases (Wasserman and Krivan 2003), this results in a low signal-to-noise ratio which complicates the immediate use of *de novo* motif detection procedures. Therefore, we developed a two-step procedure to combine the advantages of alignment and motif detection procedures.

We included a first data reduction step based on an alignment method prior to the second motif detection step (see Figure 2-1 and Methodology, §2.4.2.1). This data reduction step increases the signal-to-noise ratio in the input set used for motif detection. Data reduction is based on the assumption that longer regions conserved in the orthologs of closely related species are more likely to contain biologically relevant motifs as compared to non-conserved regions (Woolfe et al. 2005). Therefore, in our benchmark study regions, conserved among closely related orthologous

intergenic sequences of comparable size, were preselected as input for motif detection. The mammalian intergenic sequences showed a relative high overall percent identity and were comparable in length. Subsequently, these selected conserved mammalian subsequences are being subjected to motif detection, together with the full-length *Fugu* intergenic region.

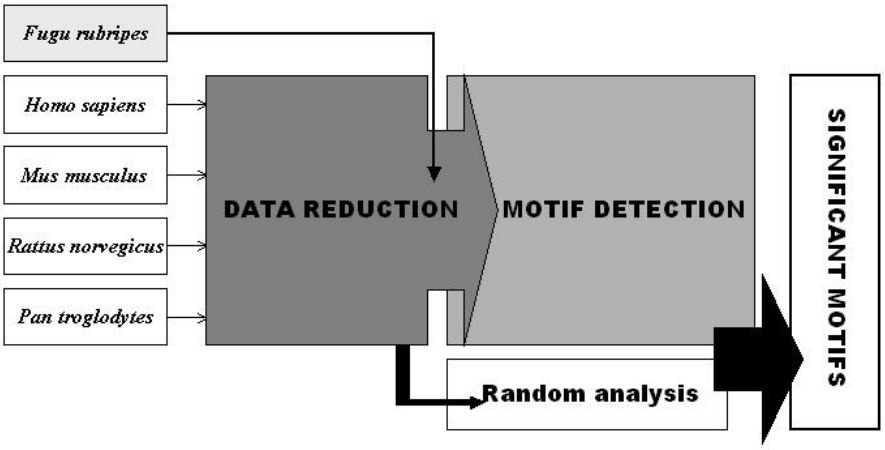


Figure 2-1. A schematic representation of the two-step developed procedure for phylogenetic footprinting. In the data reduction step, regions conserved among closely related (mammalian) orthologs are selected. Subsequently, these strongly conserved sequences are combined with a more distant ortholog (e.g., *Fugu*); this set of genes is then subjected to motif detection. Finally, significantly conserved blocks are identified using a threshold defined by a random analysis.

2.2.1.1 Data reduction

The data reduction procedure preselects subsequences conserved in closely related (mammalian) sequences. It requires a multiple alignment procedure that combines a pairwise alignment (AVID) and a clustering algorithm (TribeMCL). Details on this procedure can be found in the methodology section (§2.4.2.1). A resulting cluster consists of unique, non-overlapping subsequences, corresponding to a specific region conserved among the different related orthologs (i.e., human, chimp, mouse, and rat).

In our benchmark study, we were primarily interested in finding DNA motifs conserved among all input sequences (orthologs). Therefore, only clusters containing conserved subsequences of all mammalian orthologs included in this study (human, chimp, rat, and mouse) were retained for further analysis (see Appendix A, Table A-3).

2.2.1.2 Motif detection

The motif detection step aims at identifying motifs that are statistically overrepresented in the reduced set of orthologous intergenic sequences. To this end, we extended a previously developed Gibbs sampling based motif detection approach, MotifSampler (Thijs et al. 2001; Thijs et al. 2002a; Thijs et al. 2002b) (see Methodology, §2.4.2.2).

The adapted implementation allows the user to choose a core sequence. A potential motif is only retained when it occurs in this core sequence. Indeed, the input data for motif detection consists of a set of (mammalian) subsequences and a complete *Fugu* intergenic sequence. This *Fugu* sequence shows a relative low overall percent of identity with the other sequences. Due to the high sequence conservation (strong data dependence) between the mammalian subsequences, the original implementation of MotifSampler is not appropriate for detecting motifs in the most divergent sequence: the cost function (log likelihood score) that is optimized in the original MotifSampler offers a trade-off between the degree of conservation of the motif and the number of occurrences of the motif (Marchal et al. 2004). This results in the detection of motifs that are highly conserved between the highly similar (mammalian) sequences but that show little or no conservation with the *Fugu* intergenic sequence. Therefore, to ensure the detection of motifs conserved among all sequences, we introduced the concept of a core sequence. By selecting the most divergent ortholog (i.e., the *Fugu* sequence) as core sequence, the algorithm is forced to only detect motifs that are also present in the most distantly related organism.

The adapted implementation was also redesigned to search for long conserved blocks instead of searching for short conserved motifs only. In datasets consisting of orthologs, not only the motif itself is conserved but also the local context of the motif (Marchal et al. 2004; Woolfe et al. 2005). For this reason, we designed BlockSampler to extend motifs and search for the longest conserved blocks. A motif is thus used as a seed to generate ungapped multiple local alignments. Looking for longer motifs (i.e., blocks) also increases the specificity of motif detection (less false positives).

Finally, since it was previously shown that choosing a background model increases the performance of motif detection (Thijs et al. 2001), we adapted the algorithm such that it uses for each ortholog in the dataset an organism-specific background model.

2.2.2 Results of developed methodology on benchmark datasets

To evaluate its performance, we applied our two-step motif detection procedure to a number of benchmark datasets. Since we were primarily interested in detecting regulatory motifs over large evolutionary distances, i.e., conserved between *Fugu* and mammalian genomes, we compiled sets of evolutionary divergent vertebrate orthologs that had been described to contain conserved motifs.

In vertebrate organisms, large conserved regions tend to be associated with genes encoding regulators of development (Sandelin et al. 2004; Woolfe et al. 2005). Since our strategy aims at detecting such conserved *blocks*, we tested the methodology on three sets of orthologous genes that function in regulation of development, containing motifs described in literature: *hoxb2* (Scemama et al. 2002), *pax6* (Kammandel et al. 1999) and *scl* (Gottgens et al. 2002).

The transcription factor Hoxb2 is expressed within specific rhombomeres (i.e. a transient array of segments that compartmentalize the antero-posterior length of the hindbrain, ranging from r1 to r7) within the developing hindbrain. It also functions in the organization of the neurons in the hindbrain (Sham et al. 1993; Vieille-Grosjean et al. 1997; Davenne et al. 1999). Scemama et al. (2002) showed, using *in situ* hybridization, that the expression profile of *hoxb2* within the rhombomeres is conserved between human, mouse, zebrafish (*Danio rerio*) and striped bass (*Morone saxatilis*), whereas expression in the migrating neural crest tissues (observed in zebrafish and other vertebrates) is absent in striped bass. To assess whether the differences in expression of the respective *hoxb2* orthologs are the result of differences in regulatory elements in their corresponding intergenic regions, Scemama et al. (2002) performed a comparative analysis of the *hoxb2-hoxb3* region of striped bass, zebrafish, *Fugu*, human and mouse. This pointed out three significantly conserved regions, containing binding sites for transcription factors that are known to be responsible for *hoxb2* expression in the rhombomeres.

Pax6 is a regulatory protein that plays a crucial role in the morphogenesis of the eye. It is also an important player during the development of the brain and spinal cord and functions during the development of the pancreas. Using expression studies in mice, Kammandel et al. (1999) identified distinct elements (sequence regions) controlling tissue specific expression. Within these regulatory elements several motifs were identified that are highly conserved in the intergenic sequences of the human, mouse, and *Fugu pax6* gene (Kammandel et al. 1999).

scl encodes a transcription factor that functions in hematopoiesis and vasculogenesis. Comparative analysis of *scl* intergenic regions from human, mouse, chicken, *Fugu* and zebrafish pointed out some strongly conserved regions (Gottgens et al. 2002). In one of these regions, Göttgens et al. (2002) identified five conserved elements, three (two GATA sites and a putative SKN1) of which have previously been shown to play a role in promoter activity in hematopoietic cell lines or to be important for activity of the midbrain enhancer (Bockamp et al. 1995; Sinclair et al. 1999). Göttgens et al. (2002) showed that the two additional unnamed motifs are necessary to ensure full *scl* promoter activity in erythroid cells.

We also included in the analysis one gene, *cfos,* not related to developmental processes (Blanchette and Tompa 2002). The *cfos* gene, a member of the *fos* gene family of regulator proteins, functions in processes such as cell differentiation, proliferation and apoptosis. *fos* genes encode leucine zipper proteins that dimerize with Jun family proteins forming the AP1 complex. AP1 activity functions in a wide range of biological processes, such as cell proliferation, differentiation, apoptosis, and oncogenesis (Jochum et al. 2001). Blanchette and Tompa (2002) reported two conserved motifs in the promoters (+ 5'UTR) of *cfos* orthologs in mouse, hamster, pig, human, and *Tetraodon nigroviridis*; one of them was also conserved in the promoter region of chicken *cfos* gene. Because the *Fugu* pufferfish is closely related to the (fresh water) *Tetraodon* pufferfish, it is safe to assume that the *Fugu* intergenic sequence would also contain the two conserved motifs, detected by Blanchette and Tompa (2003) using FootPrinter (Blanchette and Tompa 2003).

All the benchmark sets consist of orthologous genes that contain evolutionary retained motifs described in literature, which have to a large extent been experimentally verified. These known motifs are used to evaluate the performance of our approach and to compare it to other algorithms. Additionally, we monitored whether our procedure was capable of detecting yet unknown motifs.

Using the two-step procedure we detected 8 significant blocks for *hoxb2*, 13 for *pax6*, 1 for *scl* and none for the *cfos* dataset (Table 2-1). The consensus scores of each of these 22 blocks are given in Table 2-2, Table 2-3 and Table 2-4 for each benchmark dataset, respectively. The location of these blocks on the complete intergenic region of the respective *Fugu* orthologs is shown in Figure 2-2; alignments can be found on the supplementary website of the Genome Biology publication (Van Hellemont et al. 2005).

Table 2-1. Conserved blocks detected in benchmark datasets. Number of blocks two-step: number of conserved blocks identified using the two-step procedure. For more details on the blocks we refer to Table 2-2 (*hoxb2*), Table 2-3 (*pax6*) and Table 2-4 (*scl*). Number of blocks UCSC: the number of blocks detected by the two-step procedure that were recovered in the USCS genome browser (aligned between mammals and *Fugu*) (Karolchik et al. 2003). Number of blocks UCR: the number of blocks detected by the two-step procedure that correspond to an ultra-conserved region (UCR) (Sandelin et al. 2004).

Gene	Number of blocks		
	two-step	UCSC	UCR
cfos	0	0	0
hoxb2	8	5	0
pax6	13	11	0
scl	1	0	0
total	22	16	0

As a first validation step, we compared our results with the alignments and conserved regions identified by well-established genome browsers, namely the UCSC genome browser (Kent et al. 2002) and the UCR browser (Sandelin et al. 2004) (Table 2-1).

The UCSC genome browser enables access to current genome assemblies; it offers visualizations of several genomic features such as cross-species homologies (Kent et al. 2002; Karolchik et al. 2003). The latter can be viewed as multiple alignments over several species, ranging from closely related mammals to more distantly related species such as chicken, zebrafish and pufferfish. The multiple alignments were generated with MULTIZ (Blanchette et al. 2004). Of the conserved 22 blocks we identified by aligning intergenic regions of mammals and *Fugu*, 16 could also be retrieved from the USCS genome browser (Table 2-1); these are indicated in Table 2-2, Table 2-3 and Table 2-4. The remaining six blocks could only be identified using our two-step approach.

The set up of the UCR browser is slightly different from the UCSC browser in that it focuses on the detection of ultra-conserved regions (UCRs) only, that is, regions conserved between human, mouse and *Fugu*. These regions were identified using sequence alignment strategies (BLAT) applied to complete genome sequences without prior data reduction (Kent 2002; Sandelin et al. 2004). Although our strategy also identifies regions highly conserved among the species under study, no overlap was detected between our conserved blocks and the UCRs (Table 2-1), that is, in the regions we studied (up to 40 kb intergenic + 5'UTR) no UCRs were located according

to the analysis of Sandelin et al. (2004). The regions UCR browser identified as ultra-conserved were located much more upstream of the gene compared to the regions we used for our analysis.

To further validate the detected blocks, we tested whether they contain the motifs that were originally reported by Scemama et al. (2002), Kammandel et al. (1999) and Göttgens et al. (2002) for *hoxb2*, *pax6* and *scl*, respectively (no significant blocks were detected for *cfos*). The previously described motifs present in the respective blocks are listed in Table 2-2, Table 2-3 and Table 2-4 (marked with an asterisk). Of the 17 motifs reported by Scemama et al. (2002), 8 were present in the significant *hoxb2*-blocks (Table 2-2). Five other motifs were present in non-significant blocks. The latter are blocks with scores that fell below the threshold we chose based on the random analysis (see Methodology, §2.4.3). The four remaining motifs could not be recovered. All motifs described by Kammandel et al. (1999) as conserved among mammalian and *Fugu pax6* intergenic regions were recovered by our methodology (Table 2-3). The conserved block detected in the *scl* dataset contains three of the five motifs previously identified by Göttgens et al. (2002) (Table 2-4); a fourth motif was picked up in a non-significant block. One motif was not detected in any of the blocks.

Besides these blocks containing known motifs, we identified several blocks (three for *hoxb2* and eight for *pax6*) that correspond to conserved regions not previously described in literature. To validate these blocks, we checked whether they were enriched for yet undescribed regulatory motifs. Hence, we screened all blocks with the TRANSFAC database of vertebrate transcription factor binding sites (Wingender et al. 2001). The result of this screening is summarized in Table 2-2, Table 2-3 and Table 2-4. As expected (Pennacchio 2003; Margulies et al. 2003) the conserved blocks we identified contain many potential binding sites; remarkably they tend to be specifically enriched for homeodomain binding sites (in blocks hoxb2 1.1, hoxb2 2.1, hoxb2 2.3, hoxb2 2.4, pax6 1.1, pax6 1.4, pax6 3.1, pax6 3.3 and scl 1.1 homeodomain binding sites were significantly overrepresented with a p-value $< 10^{-8}$). For a more detailed description of both the previously described and the new potential regulatory motifs present in the detected blocks please refer to Appendix A (§A.3.2.1).

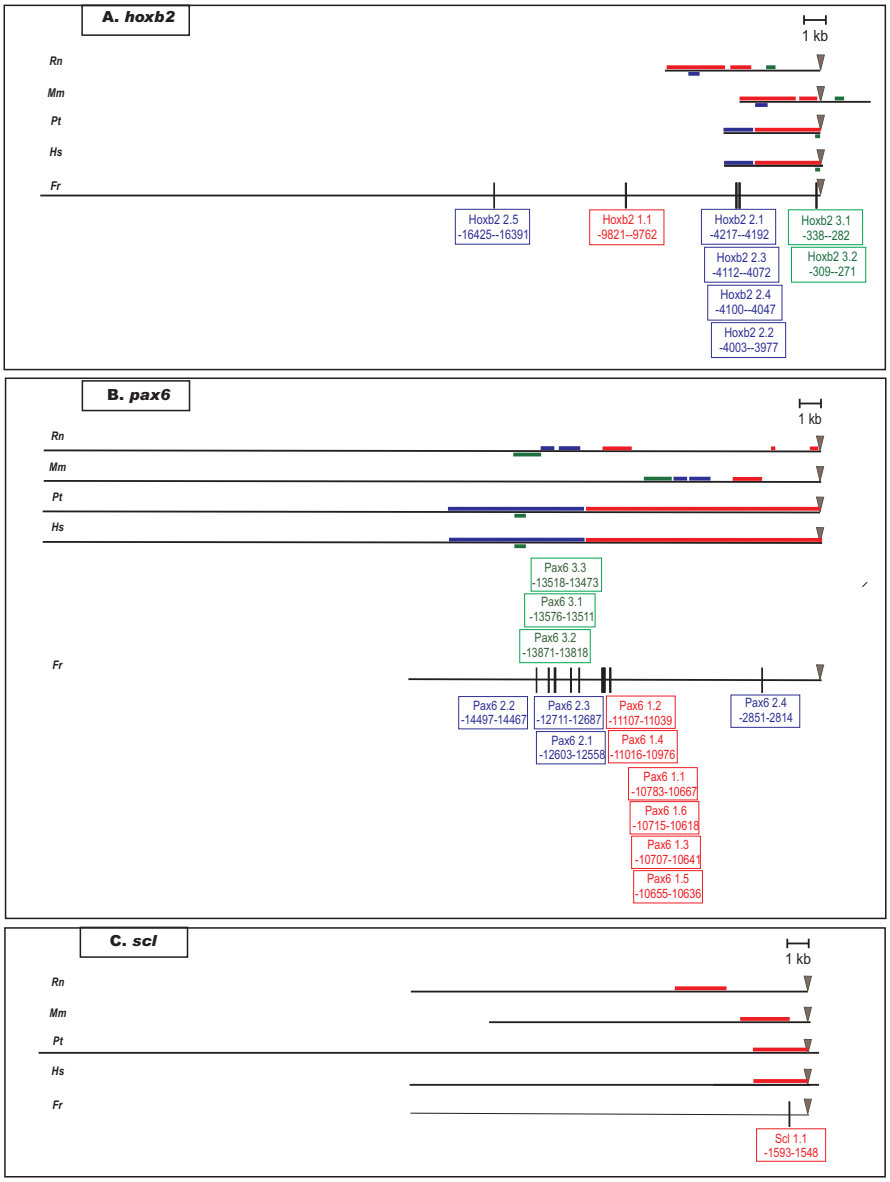


Figure 2-2. Localization of clusters and conserved blocks in the (A) *hoxb2, (B) pax6* and *(C) scl* datasets. For each dataset the different orthologous intergenic sequences are shown: *Rn: Rattus norvegicus, Mm: Mus musculus, Pt: Pan troglotydes, Hs: Homo sapiens, Fr: Fugu rubripes*. Clusters of conserved mammalian subsequences that were subjected to motif detection (that is, clusters containing at least one subsequence per mammalian organism) are represented on the respective mammalian sequences (cluster 1 in red, cluster 2 in blue en cluster 3 in green). The conserved blocks identified using BlockSampler are represented on the *Fugu* intergenic sequence (in the colour of the mammalian cluster it is located in). For each block the localization relative to the start of the *Fugu* gene is given. The transcription start sites are marked with an inverse triangle.

Table 2-2. List of the significant blocks detected in the *hoxb2* dataset. For each block, the consensus sequence is given, followed by the possible binding sites situated in this block; motifs previously described in the literature (Scemama et al. 2002), are marked with an asterisk. The motifs are summarized by their motif name (in bold), by their consensus sequence, if known, as described in the original article, by the sequence of the motif instance in our search, by the positions of the motif instance relative to the consensus sequence of the entire block and by the strand (indicated by a "+" or a "-") on which the motif occurred. Motif hits derived by TRANSFAC are indicated by their matrix accession number, the consensus of this binding site and the instances of this motif in our search. These are further characterized by their positions relative to the consensus sequence of the entire block, by the strand on which the motif occurred and by the corresponding MotifLocator score (in parentheses). The blocks identified by the UCSC genome browser as conserved between mammals and *Fugu* are marked with "UCSC", while the blocks detected by our two-step methodology but not present in the UCSC genome browser are indicated with a "-".

Block			Consensus sequence and possible binding sites
hoxb2 1.1	-		**AATTCTTTGATGCAATCGGAGGGAGCTGTCAGGGGGCTAAGATTGATCGCCTCATsTCCT**
		*	**Meis (CTGTCA),** CTGTCA 26-31 +
		*	**Hox/Pbx**, AGATTGATCG: 40-49 +
			Cap, M00253, NCANHNNN: 39-46 – (0,937); 22-29 – (0,918)
			CDP CR1, M00104, NATCGATCGS: 41-50 + (0,964)
			CDP CR3+HD, M00106, NATYGATSSS: 41-50 + (0,992)
			CdxA, M00101, AWTWMTR: 1-7 + (0,919); 6-12 + (0,903)
			HSF2, M00147, NGAANNWTCK: 40-49 + (0,925)
			MEIS1, M00419, NNNTGACAGNNN: 23-34 - (0,951)
			TGIF, M00418, AGCTGTCANNA: 24-34 + (0,966)
			Pbx1, M00096, ANCAATCAW: 39-47 – (0,909)
hoxb2 2.1	-		**TTGCACTTrGAGTTTACATTTTAATG**
		*	**octamer-motif (ATTTgCAT),** GTTTACAT: 12-19 +
		*	**Adhf-2a (TGCACTgAGA),** TGCACTTrGA: 2-11 +
			CdxA, M00101, AWTWMTR: 20-26 + (0,978); 19-25 - (0,905); 17-23 - (0,927)
			SRY, M00148, AAACWAM: 14-20 - (0,905)
hoxb2 2.2	UCSC		**AAAAnTGTACTTTTTTAGTATTTACyT**
		*	**HoxA5 (TTTAaTAaTTA),** TTTAGTATTTA: 14-24 +
			CdxA, M00101, AWTWMTR: 16-22 - (0,979)
			SRY, M00148, AAACWAM: 7-13 - (0,928)
hoxb2 2.3	UCSC		**GTGTGTTCTAGTGAACATTTTCATATATATTTATTGGTTAT**
		*	**glucocorticoid receptor,** AGTGAACA: 10-17 +
		*	**CCAAT BOX,** ATTGGTT: 27-33 +
			Cap, M00253, NCANHNNN: 15-22 + (0,919); 21-28 + (0,906); 7-14 - (0,919)
			CdxA, M00101, AWTWMTR: 23-29 + (0,958); 29-35 + (0,940); 28-34 - (0,956); 26-32 - (0,951); 24-30 -(0,958); 22-28 - (0,960)
			FOXJ2, M00422, NNNWAAAYAAAYANNNNN: 23-40 - (0,932)

			HFH-3, M00289, KNNTRTTTRTTTA: 25-37 + (0,908)
			NF-Y, M00185, TRRCCAATSRN: 30-40 - (0,914)
			Oct-1, M00162, CWNAWTKWSATRYN: 14-27 + (0,913)
			Pbx-1, M00096, ANCAATCAW: 30-38 - (0,948)
hoxb2 2.4	UCSC		**GTGAACATTTTCATATATATTTATTGGTTATAGCCTGTTAAAATATTTTCTTTT**
		*	**GATA 1,** TTATAGCC: 28-35 +
		*	**CCAAT BOX**, ATTGGTT: 23-29 +
			Cap, M00253, NCANHNNN: 5-12 + (0,919); 11-18 + (0,906)
			CCAAT box, M00254, NNNRRCCAATSA: 21-32 - (0,940)
			CdxA, M00101, AWTWMTR: 13-19 + (0,958); 19-25 + (0,940); 39-45 + (0,925); 46-52 + (0,901); 36-42 - (0,930); 18-24 -(0,957); 16-22 - (0,951); 14-20 - (0,958); 12-18 - (0,960)
			FOXD3, M00130, NAWTGTTTRTTT: 41-52 + (0,924)
			FOXJ2, M00422, NNNWAAAYAAAYANNNNN: 13-30 - (0,932)
			HFH-3, M00289, KNNTRTTTRTTTA: 15-27 + (0,908)
			HNF-3beta, M00131, KGNANTRTTTRYTTW: 39-53 + (0,920)
			NF-Y, M00185, TRRCCAATSRN: 20-30 - (0,914)
			Oct-1, M00162, CWNAWTKWSATRYN: 4-17 + (0,913)
			Pbx-1, M00096, ANCAATCAW: 20-28 - (0,948)
			SRY, M00148, AAACWAM: 47-53 - (0,961)
hoxb2 2.5	UCSC		**AATTCyCTCTTGGAACTTTCTTTGTTCTTCmGTAG**
			HSF1 , M00146, AGAANRTTCN: 12-21 + (0,915); 12-21 - (0,930)
			HSF2, M00147, NGAANNWTCK: 12-21 + (0,948); 12-21 - (0,930)
			SRY, M00148, AAACWAM: 17-23 - (0,961)
hoxb2 3.1	UCSC		**GGCCnAGACnAGCGATTGGCGGAGrCCGGTCCCGTGACCAnGAATTCCCTGyAATTT**
			NF-Y, M00185, TRRCCAATSRN: 12-22 - (0,915)
			USF, M00187, CYCACGTGNC: 29-38 - (0,957)
			USF, M00217, NCACGTGN: 30-37 + (0,902)
hoxb2 3.2	-		**TCCCGTGACCAnGAATTCCCTGyAATTTCGnyGGAGTCC**
			USF, M00217, NCACGTGN: 1-8 + (0,902)

Table 2-3. List of the significant blocks detected in the *pax6* dataset. For legend see Table 2-2.

Block			Consensus sequence and possible binding sites
pax6 1.1	UCSC		**CTTAATGATGAGAGATCTTTCCGCTCATTGCCCATTCAAATACAATTGTAGATCGAAGCCGGCCTT**
			GTCAsGTTGAGAAAAAGTGAATTTCTAACATCCAGGACGTGCCTGTCTACT
		*	**Minimal fragment for expression in lens and cornea as described in *: 11-117 +**
			Cap, M00253, NCANHNNN: 25-32 + (0,940); 79-86 - (0,964); 4-11 - (0,946); 1-8 - (0,903)
			CCAAT box, M00254, NNNRRCCAATSA: 27-38 + (0,901)
		*	**CdxA**, M00100, "MTTTATR": 1-7 + (0,921) *; 87-93 + (0,913)
		*	**CdxA**, M00101, AWTWMTR: 1-7 + (0,934); 4-10 + (0,921); 38-44 + (0,905), 87-93 + (0,988)
			c-Ets-1(p54), M00032, NCMGGAWGYN: 98-107 + (0,906)
			c-Ets-1(p54), M00074, NNACMGGAWRTNN: 92-104 - (0,901)
			En-1, M00396, GTANTNN: 37-43 - (0,967)
			GATA-3, M00351, ANAGATMWWA: 11-20 + (0,920)
			HSF2, M00147, NGAANNWTCK: 13-22 - (0,933)
			p53, M00272, NGRCWTGYCY: 101-110 + (0,949)
pax6 1.2	UCSC		**CATTATTGTTGCCAGCACGAAGCATCACAATCAATCATAAGGAAGTCCAGTTGGCAGGTGTCAATCTTG**
			CdxA, M00101, AWTWMTR:1-7 - (0,995)
			Cap, M00253, NCANHNNN: 25-32 + (0,934), ; 31-38 + (0,903); 35-42 + (0,903); 47-54 + (0,908); 61-68 + (0,937)
			CDP CR3+HD, M00106, NATYGATSSS: 27-36 - (0,907)
			c-Ets-1(p54), M00074, NNACMGGAWRTNN:36-48 + (0,902)
		*	**HOXA3**, M00395, CNTANNNKN: 1-9 + (0,905)
			MyoD, M00184, NNCACCTGNY: 53-62 - (0,956)
		*	**Pbx-1**, M00096, ANCAATCAW: 30-38 + (0,986); 2-10 - (0,923)
			Sox-5, M00042, NNAACAATNN: 3-12 - (0,932)
			SRY, M00148, AAACWAM: 33-39 + (0,910)
			USF, M00122, NNRNCACGTGNYNN: 51-64 + (0,913); 51-64 - (0,908)
pax6 1.3	UCSC		**GAAAAAGTGAATTTCTAACATCCAGGACGTGCCTGTCTACTTTCAGwGAATTGCATCCAATCACCCC**
			Cap, M00253, NCANHNNN:3-10 - 0,964
			CCAAT box, M00254, NNNRRCCAATSA: 52-63 + (0,949)
			CdxA, M00100, "MTTTATR":11-17 + (0,913)
			CdxA, M00101, AWTWMTR:11-17 + (0,988)
			c-Ets-1(p54), M00032, NCMGGAWGYN: 22-31 + (0,906)
			c-Ets-1(p54), M00074, NNACMGGAWRTNN:16-28 - (0,901)
			En-1, M00396, GTANTNN: 58-64 - (0,948)
			GATA-1, M00075, SNNGATNNNN: 56-65 - (0,930)
			GATA-3, M00077, NNGATARNG: 56-64 - (0,917)
			NF-Y, M00185, TRRCCAATSRN: 54-64 + (0,910)
			p53, M00272, NGRCWTGYCY:25-34 + (0,949)
			SRY, M00148, AAACWAM: 59-65 + (0,917)
pax6 1.4	UCSC		**GTCTATATTTAATCCAATTATAAGGGTCACGGAGTAAGTGC**

	*	**Motif containing homeoboxes described by Kammandel et al. (1999)** TTTAATCCAATTATAA: 8-23 +
		Cap, M00253, NCANHNNN: 34-41 - (0,904)
		CdxA, M00100, "MTTTATR":16-22 + (0,907)
		CdxA, M00101, AWTWMTR: 16-22 + (0,995); 16-22 - (0,906); 6-12 - (0,931); 4-10 - (0,951)
		En-1, M00396, GTANTNN:15-21 - (0,948)
		Nkx2-5, M00240, TYAAGTG: 34-40 + (0,927)
		RORalpha1, M00156, NWAWNNAGGTCAN: 18-30 + (0,919)
		TCF11, M00285, GTCATNNWNNNNN: 26-38 + (0,906)
pax6 1.5	UCSC	**GCATCCAATCACCCCCAGGG**
		Cap, M00253, NCANHNNN: 9-16 + (0,965)
		En-1, M00396, GTANTNN: 6-12 - (0,948)
		GATA-3, M00077, NNGATARNG: 4-12 - (0,917)
		SRY, M00148, AAACWAM:7-13 + (0,917)
pax6 1.6	UCSC	**CAsGTTGAGAAAAAGTGAATTTCTAACATCCAGGACGTGCCTGTCTACTTTCAGw**
		GAATTGCATCCAATCACCCCCAGGGAATTCnGCTAATGTCTCC
	*	**Homeobox-binding site described by Kammandel et al. (1999)**, GCTAATGTCTC: 87-97 +
		Cap, M00253, NCANHNNN: 69-76 + (0,965); 87-94 – (0,903); 11-18 – (0,964)
		CCAAT box, M00254, NNNRRCCAATSA: 60-71 + (0,949)
		CdxA, M00100, "MTTTATR":19-25 + (0,913)
		CdxA, M00101, AWTWMTR: 19-25 + (0,988)
		c-Ets-1(p54), M00032, NCMGGAWGYN: 30-39 + (0,906)
		c-Ets-1(p54), M00074, NNACMGGAWRTNN: 24-36 – (0,901)
		En-1, M00396, GTANTNN: 66-72 – (0,948)
		GATA-1, M00075, SNNGATNNNN: 64-73 – (0,930)
		GATA-3, M00077, NNGATARNG: 64-72 – (0,917)
		NF-Y, M00185, TRRCCAATSRN: 62-72 + (0,910)
		p53, M00272, NGRCWTGYCY: 33-42 + (0.949)
		SRY, M00148, AAACWAM: 67-73 + (0,917)
pax6 2.1	UCSC	**TGGGTCCATTTTCCAGAyGGTTTGTTACTCTTGCTGCmTGATTTrG**
		Cap, M00253, NCANHNNN:6-13 + (0,921)
		CdxA, M00101, AWTWMTR:9-15 + (0,918)
		SRY, M00148, AAACWAM: 21-27 - (0,942)
pax6 2.2	-	**ATTTTGGTTGCTTTCAGGTwTAATTAACTTT**
		Nkx2-5, M00241, CWTAATTG: 21-28 - (0,902)
pax6 2.3	UCSC	**ATTGTAATCATTTCAATTATCTTCA**
		Cap, M00253, NCANHNNN: 8-15 + (0,927)
		En-1, M00396, GTANTNN: 14-20 - (0,948)
		Nkx2-5, M00241, CWTAATTG: 14-21 - (0,930)
pax6 2.4	-	**GGTTGCTTTCAGGTwTAATTAACTTTGAACAACAAATA**
		Nkx2-5, M00241, CWTAATTG:16-23 - (0,902)

pax6 3.1	UCSC	**TTGTAATTACTGCCCTTCATGTGGTCCGGTGCCTTGAACCATCTTTAATTAAAAGCATAATTAAGG**
		AML-1a, M00271, TGTGGT: 20-25 + (1,000)
		Cap, M00253, NCANHNNN: 39-46 + (0,910); 55-62 + (0,909); 6-13 - (0916)
		CdxA, M00100, MTTTATR: 56-62 - (0,934)
		CdxA, M00101, AWTWMTR: 6-12 + (0,988); 44-50 + (0,913); 47-53 + (0,900); 48-54 + (0,905); 59-65 + (0,903); 60-66 + (0,926); 56-62 - (0,998); 47-53 - (0,913); 44-50 - (0,901); 43-49 - (0,907); 2-8 - (0,949);
		En-1, M00396, GTANTNN: 3-9 + (0,912); 4-10 - (0,912)
		HSF2 , M00147, NGAANNWTCK: 35-44 + (0,908)
		Nkx2-5, M00241, CWTAATTG: 56-63 + (0,935), 58-65 - (0,954)
		USF, M00217, NCACGTGN: 17-24 - (0,921)
pax6 3.2	UCSC	**AAGGCTTGCAGCTGCCTCCAAATCAATAGAyGTCAAAGAAATATGAAAACArTC**
		CdxA, M00101, AWTWMTR: 39-45 + (0,953); 36-42 - (0,925)
		SRY, M00148, AAACWAM: 35-41 + (0,961)
		Cap, M00253, NCANHNNN: 8-15 + (0,931); 39-46 - (0,940); 8-15 - (0,931)
		AP-4 , M00175, VDCAGCTGNN: 7-16 - (0,902)
		MyoD, M00184, NNCACCTGNY: 7-16 + (0,957)
		SRY, M00160, NWWAACAAWANN: 19-30 + (0,928)
pax6 3.3	UCSC	**GCATAATTAAGGGAAGATCTAAAGAAAGACAATTACCAGATGGTCT**
		Cap, M00253, NCANHNNN: 1-8 + (0,909)
		CdxA, M00100, MTTTATR: 2-8 - (0,934)
		CdxA, M00101, AWTWMTR: 5-11 + (0,903); 6-12 + (0,926); 32-38 + (0,939); 2-8 - (0,998)
		En-1, M00396, GTANTNN: 30-36 - (1,000)
		GATA-1, M00075, SNNGATNNNN: 36-45 + (0,936)
		GATA-2, M00076, NNNGATRNNN: 36-45 + (0,922)
		GATA-3, M00351, ANAGATMWWA: 13-22 + (0,949)
		HOXA3, M00395, CNTANNNKN: 29-37 - (0,939)
		Msx-1, M00394, CNGTAWNTG: 30-38 - (0,915)
		MyoD, M00184, NNCACCTGNY: 35-44 - (0,919)
		Nkx2-5, M00241, CWTAATTG: 2-9 + (0,935); 4-11 - (0,954)
		SRY, M00148, AAACWAM: 21-27 + (0,961); 25-31 + (0,927)
		USF, M00122, NNRNCACGTGNYNN: 33-46 + (0,907); 33-46 - (0,904)

Table 2-4. List of the significant blocks detected in the *scl* dataset. For legend see Table 2-2.

Block			Consensus sequence and possible binding sites
scl 1.1	-		**TTGCCAAATTAAAATGAATCATTTGGCCCATAATGGCCGAGGCGCT**
		*	**Conserved sequence identified by Göttgens et al (2002)**, GCCAAAT: 3-9 +
		*	**Putative SKN1 site reported by Göttgens et al (2002)** AATGAATCATTT: 13-24 +
			CdxA, M00100, "MTTTATR": 29-35 - (0,917)
			CdxA, M00101, AWTWMTR: 7-13 + (0,901), 8-14 + (0,905); 10-16 + (0,927); 29-35 + (0,927); 29-35 - (0,929); 7-13 - (0,913)
		*	**En-1**, M00396, GTANTNN: 30-36 + (0,936)
			Cap, M00253, NCANHNNN: 19-26 + (0,932); 10-17 - (0,933)
			Pbx-1, M00096, ANCAATCAW:14-22 + (0,941)
			AP-1, M00199, NTGASTCAG: 14-22 + (0,913)
		*	**HOXA3**, M00395, CNTANNNKN: 29-37 + (0,927)
			Tst-1, M00133, NNKGAATTAVAVTDN: 3-17 + (0,901)

Besides these well-described benchmark datasets, we applied our method to six additional datasets, differing in composition from the benchmark datasets (see Appendix A, §A.2: Table A-2). They all contained a combination of four mammalian sequences (rat, mouse, human, chimp or dog) to be used in the data reduction step and an additional set of sequences originating from more distantly related orthologs (chicken, *Fugu*, *Tetraodon nigroviridis* and zebrafish in different combinations) added in the motif detection step. Four of the six additional datasets are derived from genes functioning in developmental regulation including three homeobox genes (*ghs1*, *Meis2*, *hoxb5*) and one zinc finger protein *egr3*. Besides these regulators involved in development, two genes, *pcdh8* and *hiv-ep1* were included which are, according to our knowledge, unrelated to development. *pcdh8* is believed to function as a calcium-dependent cell-adhesion protein and *hiv-ep1* binds to enhancer elements present in several viral promoters and in a number of cellular promoters such as those of the class I MHC, interleukin-2 receptor, and interferon-beta genes. In the additional datasets involved in development, we detected several strongly conserved blocks (see Appendix A, §A.3.2.2: Table A-5): *ghs1* contained four blocks that are conserved among human, chimp, mouse, rat, and pufferfish (*Fugu* and *Tetraodon)*; in *meis2* two blocks were recovered that are retained in all organisms under study except for *Fugu;* and in *hoxb5* six strongly conserved blocks were detected in mammals and pufferfish, while the motif seems to have been lost in chicken. In *egr3,* two blocks were found conserved in mammals and fish. In the non-developmental related datasets only in *pcdh8* one large block was detected, conserved in human, chimp, mouse, rat, chicken, *Tetraodon* and *Fugu,* but not in zebrafish. This shows that conserved regions might also exist in genes not involved in development,

although a possible involvement of this additional gene in developmental processes cannot be ruled out. Detailed results of these analyses can be found in Appendix A (Tables A-6 to A-10). Because, in contrast to the benchmark datasets, the motifs in these additional datasets have not been studied so extensively as those of the benchmarks, we cannot guarantee all detected blocks are biologically functional.

2.2.3 Evaluation of the developed procedure

To compare the performance of our newly developed two-step strategy to that of other frequently used algorithms, we evaluated to what extent MotifSampler (Thijs et al. 2002a), MAVID (Bray and Pachter 2004) and 'Threaded Blockset Aligner' (TBA) (Blanchette et al. 2004) could recover known motifs in our benchmark sets.

Table 2-5. Comparison of two-step procedure with other methodologies. # motifs: the number of motifs reported by Blanchette and Tompa (2002) in *cfos*, Scemama et al. (2002) in *hoxb2*, Kammandel et al. (1999) in *pax6* and Göttgens et al. (2002) in *scl*. Two-step BS: the number of previously described motifs detected by two-step procedure, combining data reduction and motif detection using BlockSampler. The numbers in parentheses are the number of motifs present in non-significant blocks. BS: the number of previously described motifs detected by BlockSampler in initial full-length datasets. Two-step MS: the number of previously described motifs detected by combining data reduction and motif detection using MotifSampler. MS: the number of previously described motifs detected by MotifSampler in initial full-length datasets. MAVID: the number of previously described motifs detected (correctly aligned) by MAVID. TBA: the number of previously described motifs detected by TBA. * Only part of a motif was detected.

Gene	# motifs	Two-step BS	BS	Two-step MS	MS	MAVID	TBA
cfos	2	0	0	0	0	0	0
hoxb2	17	8 (+5)	13	2	1	0	0
pax6	6	6	1*	0	0	6	6
scl	5	3 (+1)	1	0	0	0	0
total	30	17 (+6)	15	2	1	6	6

First, we studied the performance of the alignment algorithms MAVID and TBA in detecting conserved regions within our four benchmark datasets. Since MAVID and TBA were originally developed to perform multiple alignments on long sequences, we applied these algorithms to the initial full-length benchmark datasets, i.e. complete mammalian and *Fugu* intergenics. We evaluated to what extent motifs or conserved regions described in original articles were correctly aligned using either MAVID or TBA. The results are summarized in Table 2-5 (MAVID and TBA columns) and in Appendix A (§A.4.1).

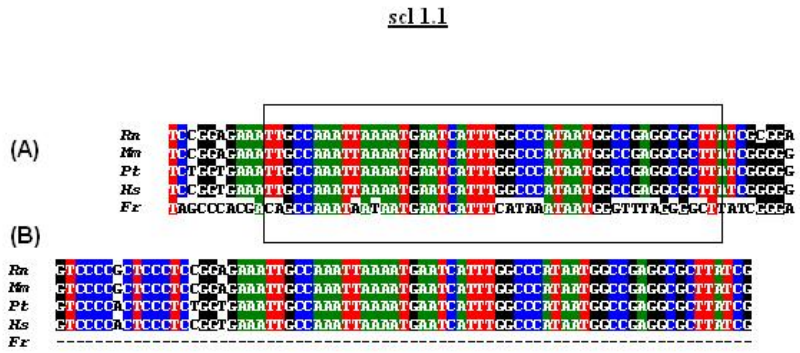


Figure 2-3. Comparison of two-step strategy with MAVID for *scl* dataset. (A) Conserved block: alignment of the different *scl* orthologs. The conserved block as identified by BlockSampler is marked with a boxed area. (B) Visualization of the MAVID alignment of the corresponding region. The dashed line denotes a gap in the alignment. *Rn: Rattus norvegicus, Mm: Mus musculus, Pt: Pan troglotydes, Hs: Homo sapiens, Fr: Fugu rubripes.*

MAVID alignment of all three *cfos* datasets (i.e. mammalian orthologs combined with each of the three *Fugu* paralogs) could not recover either of the two motifs previously described by Blanchette and Tompa (2002) (Table 2-5). This is in line with our results showing the overall low homology between the *cfos* mammalian and *Fugu* orthologs. The MAVID-alignment of most of the *hoxb2* blocks containing previously described motifs shows that a conserved region in the mammalian intergenic sequences is broken up in small conserved parts interrupted by gaps when aligned to the longer *Fugu* sequence, resulting in an incorrect alignment of the regulatory motifs: previously reported motifs were not recovered in the MAVID alignment (Table 2-5 and Appendix A: Figure A-1). Our method performs better because the most heterogeneous sequence is only aligned in a second step, using a high flexible local alignment procedure (BlockSampler). Regarding *pax6,* most of the blocks containing previously described motifs were correctly aligned by MAVID and all the motifs described by Kammandel et al. (1999) could be correctly retrieved over all

the orthologs under study (Table 2-5 and Appendix A: Figure A-2). This dataset is probably relatively well suited for MAVID because the mammalian sequences are only twice as large as the pufferfish *pax6* intergenic region (Table 2-6). Although the lengths of the intergenic regions in the *scl* dataset (Table 2-6) are in the same order of magnitude (ranging from 16,5 to 40kb), MAVID did not succeed in identifying any of the motifs previously described by Göttgens et al. (2002) (Figure 2-3, Table 2-5).

Although TBA has been shown to outperform MAVID in aligning more divergent sequences (Blanchette et al. 2004), applying this alignment tool to the benchmark datasets generated similar results as MAVID: all known *pax6*-regulating motifs were detected, while motifs present in the other benchmark datasets were not recovered (Table 2-5, TBA column).

Besides detecting the blocks with previously described motifs, our two-step methodology also discovered new blocks (block pax6 2.4, for instance) that could not be recovered when aligning the intergenic sequences with MAVID or TBA (Ureta-Vidal et al. 2003) (see Appendix A, §A.4.2).

Overall, based on our benchmark analysis, the two-step method performs better than MAVID or TBA in identifying conserved blocks in distantly related orthologs: the proposed method is able to recover in our benchmark sets all the known motifs identified by MAVID and TBA but in addition finds a number of previously described motifs ignored by these algorithms (Table 2-5, two-step BS, MAVID and TBA columns). Using the two-step procedure, first selecting strongly conserved orthologous sequences, clearly facilitates alignment with the more divergent (lower overall similarity) sequence.

We also tested the performance of MotifSampler, as an example of a probabilistic motif detection procedure on the unreduced dataset. In this case, only one previously described motif was detected (Table 2-5, MS column). This was to be expected as in unreduced datasets the signal-to-noise ratio is too high for standard motif detection procedures to give reliable and interpretable results.

Our two-step procedure includes two adaptations over previous existing methods: first, it allows for a data reduction step, secondly, we developed a motif detection procedure specifically adapted to the purpose of detecting large conserved blocks (BlockSampler). To assess the relative contribution of each of these adaptations to the overall result, we set up the following experiment: to study the specific influence of the data reduction step, we compared the results of applying BlockSampler to both the unreduced benchmark datasets and the datasets obtained after data reduction. Table 2-5 (BS and two-step BS columns) shows the results of this comparison. Overall, the results seem comparable: application of BlockSampler to the complete intergenic sequences results in recovery of 15

of the 30 previously reported motifs (in all four datasets), while the two-step method identified 17. Thus, at first sight, there does not seem to be a major contribution from the data reduction step. A closer look at Table 2-5, however, shows that the positive contribution of the data reduction (increasing signal-to-noise ratio) is strongly dependent on the lengths of the intergenic sequences to be aligned. A major positive effect is observed for the large *pax6* and *scl* datasets, whereas for the *hoxb2* set of which the sequences under study are rather short, the data reduction does not offer a clear advantage. To assess the specific improvements of using BlockSampler instead of standard motif detection approaches, we compared the results of the BlockSampler to those of the MotifSampler when both were applied to the reduced datasets. A reduced dataset thus consists of a subcluster of mammalian sequences (Figure 2-4) and a complete *Fugu* ortholog. The performance of MotifSampler was far below that of BlockSampler: MotifSampler only detected two previously described motifs (Table 2-5, two-step MS column), both in the *hoxb2* set, while BlockSampler recovered 17 previously described motifs (Table 2-5, two-step BS column). Moreover because MotifSampler searches for short motifs (default 8 nt), it detects many false positive hits. These results show that independent of the data reduction step, BlockSampler is clearly more suited for detecting large conserved blocks than MotifSampler.

2.3 Discussion

We developed a two-step methodology to search for regions (motifs) conserved over different phylogenetic lineages in long intergenic sequences of heterogeneous size. In a first step, an alignment method is used to select conserved subsequences in intergenic orthologous sequences of comparable size of closely related vertebrate genomes, since these are expected to be enriched for regulatory motifs (Pennacchio 2003; Woolfe et al. 2005). The combination of this preselected dataset of conserved sequences and the full-length intergenic sequence of a more distant ortholog, which therefore is more likely to differ in size and overall homology, is subjected to probabilistic motif detection. The preselection step facilitates motif detection by enhancing the signal-to-noise ratio in the dataset. For the second motif detection step we used an extension of a Gibbs sampling based algorithm (Thijs et al. 2002a) with a higher performance in detecting large conserved blocks within a set of orthologous sequences. Using the strategy mentioned above we could combine the advantages of alignment methods, which have been shown to be very suitable for aligning long highly conserved intergenic sequences and of the probabilistic algorithms for motif detection that usually are more appropriate when looking for smaller regions of conservation (lower degree of similarity).

We applied this two-step methodology to four well-studied datasets for which functional phylogenetically conserved motifs had extensively been described. Our approach identified most of the previously described motifs. In addition, we detected several blocks, not previously described in literature or not present in any of the two genome browsers (UCSC, UCR) we compared our results with. Because highly conserved blocks most probably consist of consecutive transcription factor binding sites (Pennacchio 2003; Margulies et al. 2003; Woolfe et al. 2005), we screened the conserved blocks, with the TRANSFAC motif database (Wingender et al. 2001). These blocks contained abundant copies of homeodomain binding sites. This is not unexpected since most of the genes we were studying function in regulation of development (Boffelli et al. 2004; Woolfe et al. 2005). These blocks most probably contain, besides the motifs obtained with the TRANSFAC screening, many more motifs not yet annotated in TRANSFAC. Alternatively, they might have other, not yet characterized biological functions, e.g. transcripts of unknown function (TUFs) (ENCODE project).

Some previously described motifs were missed, however, because of the strong selection criteria we used: since regulatory elements tend to be grouped (Pennacchio 2003; Thomas et al. 2003; Margulies et al. 2003; Woolfe et al. 2005), we assumed that the sequences surrounding a regulatory motif are also conserved (due to the presence of other binding sites). Motifs located in a variable context will probably go undetected.

By applying our method to additional datasets, with a configuration different from the benchmark dataset we could demonstrate that our methodology is more generally applicable.

Comparing the performance of the two-step procedure with that of MAVID and TBA, as representatives of multiple alignment methods and with the performance of MotifSampler, as an example of a motif detection method, showed that our approach outperformed these alternative methods when the intergenic sequences became either too long or too heterogeneous in size.

Additionally, we studied the marginal contribution of the data reduction step and the improved method for motif detection on the final performance of the two-step procedure: BlockSampler performed on overall better than the related algorithm MotifSampler, both on long sequences and on intergenic regions reduced in size. The data reduction step seemed essential when the length of the intergenic sequences to be compared becomes excessive.

Although our two-step procedure has proven successful, there still is room for improvement, for instance by taking into account the phylogenetic relationships between the sequences under study in the second motif detection step. The contribution of finding a motif in an ortholog to the

global motif score could be weighted according to its phylogenetic distance with the other sequences in which the motif is also present. Indeed, this way we would account for the specific composition of a dataset because closely related orthologs are less informative than further related ones. Other adaptations that can be made: if one wants to relax the assumption of conserved order of motifs in the first data reduction step, it suffices to replace AVID in this step by a more local aligner such as BLAT (Kent 2002). Also our motif detection algorithm could be extended for more advanced background models (Down and Hubbard 2005).

2.4 Methodology

2.4.1 Benchmark datasets

The benchmark datasets were generated as follows: first, a set of orthologous genes was defined using the Ensembl genome browser version 23 (2005b). In this study, the benchmark datasets included genes from human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), chimp (*Pan troglodytes*), and pufferfish (*Fugu rubripes)* (Appendix A: Table A-1). Regarding the *cfos* dataset, Ensembl identified three *Fugu* paralogs - SINFRUG00000132418, SINFRUG00000132419 and SINFRUG00000143787 - that were all included in the analysis. The additional datasets *erg3*, *gsh1*, *hiv-ep1*, *hoxb5*, *meis2* and *pcdh8* contain multiple distantly related orthologs (Appendix A, §A.2: Table A-2).

Table 2-6. Base pair lengths of the intergenic sequences for each benchmark dataset. The *Fugu cfos* intergenic sequences are derived from *SINFRUG00000132418, †SINFRUG00000132419 and ‡SINFRUG00000143787. The Ensembl-ids (+ 1 Genbank accession number) are given in Table A-1 (Appendix A). *Hs: Homo sapiens, Mm: Mus musculus, Rn: Rattus norvegicus, Pt: Pan troglotydes, Fr: Fugu rubripes*.

Gene	*Hs*	*Mm*	*Rn*	*Pt*	*Fr*
cfos	40154	33157	40132	40154	3606*
					3606†
					1244‡
hoxb2	4973	6744	7640	4878	39219
pax6	40102	40000	40000	40000	21204
scl	20981	16471	20343	39999	20155

Subsequently, the intergenic regions of these orthologs were selected using the Ensembl mart database release 21.1. The region upstream of the transcription start (as defined by Ensembl) was limited to 40 kb. Additionally, the 5'UTR was included. Lengths of the respective intergenics are given in Table 2-6; the benchmark datasets, *cfos*, *hoxb2*, *pax6* and *scl* can be found as supplementary information. The rat *cfos* ortholog ENSRNOG00000008015, *Fugu hoxb2* ortholog SINFRUG00000136637, chimp *pax6* ortholog ENSPTRG00000003474, and *scl* chimp ENSPTRG00000003474 contain long N-stretches, probably as a result of incomplete preliminary annotation.

Remarkably, where *Fugu* is known to have a very compact genome (Brenner et al. 1993), the *Fugu hoxb2* mentioned above is very long as compared to the mammalian *hoxb2* intergenic sequences (Table 2-6). This is probably due to the presence of a pseudogene (SINFRUG00000157209) in the intergenic region of SINFRUG00000136637 at circa 5.9 kb from the transcription start site of *hoxb2* which was not yet annotated in the release version 23 of Ensembl. This is explained in more detail in Appendix A (§A.3.2.1).

All intergenic sequences were selected as described above, except the intergenic sequence of the *Fugu scl* ortholog. Because the putative *scl* ortholog annotated by Ensembl (SINFRUG00000145588) did not contain motifs shown to be present in the *Fugu scl* ortholog by Göttgens et al. (2002), we used the Genbank *Fugu scl* sequence [Genbank: AJ131019]. This sequence, (referring to a cosmid sequence of circa 33 kb) was also used in the original study of Barton et al. (2001) . In order to delineate the intergenic region of *scl*, we aligned the coding sequence from the *scl* homolog SINFRUG00000145588 with the AJ131019 sequence using 'blast 2 sequences' (Tatusova and Madden 1999). The coding region was located from positions 20156 to 22165; we then selected the upstream region (from positions 1 to 20155).

2.4.2 A two-step procedure for phylogenetic footprinting

A schematic representation of the developed two-step procedure is given in Figure 2-1.

2.4.2.1 Step I: Data reduction

In this step, a dataset consisting of the complete intergenic sequences of comparable size originating from orthologs of closely related organisms is reduced to a dataset of preselected sequences conserved among

all/most compared orthologs. First, related vertebrate intergenic regions of comparable size (in this study these sequences correspond to the mammalian sequences, (i.e. human, chimp, rat, and mouse)) are aligned using the pairwise alignment algorithm AVID (using default parameters) (Bray et al. 2003). For each ortholog, sequences corresponding to the significantly conserved regions of the pairwise alignment are selected using VISTA (Frazer et al. 2004). Significance of the alignment is defined by two parameters (VISTA parameters): the window length (L), the region for which the percent identity is calculated; and the conservation level (C) in the selected window, the minimal percent identity of the aligned region to be considered as significantly conserved. The parameter settings were adapted to the evolutionary distance of the compared organisms. The closer the organisms were related, the higher the threshold on the degree of conservation was chosen. The conservation parameters used were: for human-mouse comparison 85% over 200 nt, human-rat 85% over 200 nt, human-chimp 85% over 350 nt, human-dog 80% over 200 nt, mouse-rat 85% over 350 nt, mouse-chimp 85% over 200 nt, mouse-dog 80% over 200 nt and for rat-chimp 85% over 200 nt, rat-dog 80% over 200 nt and chimp-dog 80% over 200 nt (see also Appendix A, §A.5.1 on how these parameters were set).

To identify orthologous regions conserved in multiple related vertebrate sequences of comparable size (that is, multiple alignment), homologies between all preselected sequences were determined (using AVID with default parameters). Subsequently, multiple conserved regions were identified using the graph based clustering TribeMCL (Enright et al. 2002). We choose for TribeMCL as this is a well-known graph-based clustering algorithm that was originally designed to recover transitivity relations between biological sequences (i.e., orthologous proteins). Each resulting cluster corresponds to a region conserved in multiple sequences and consists of a set of preselected sequences, originating from the different related orthologs of comparable size that mutually show a minimal degree of conservation. Several runs of TribeMCL were performed for each dataset, using different values of clustering parameters I and P (for more details we refer to Appendix A, §A.5.2). The parameter I did not seem to have a major influence on the size of the clusters and, therefore, was set at 4. For the P value three different values were tested per dataset and the parameter that resulted in small tightly linked clusters was chosen as these clusters correspond to strongly conserved regions. The parameters of choice for the benchmark datasets were: for *cfos,* I=4 and P=0; for *hoxb2,* I=4 and P=-10; for *pax6,* I=4 and P=0; and for *scl,* I=4 and P=-10. Concerning the additional datasets, the parameter setting of choice was I=4 and P=0 for *egr3*, *hiv-ep1*, hoxb5, *meis2* and *pchd8* and I=4 and P=-10 for *ghs1*.

Some clusters contain different subsequences derived from the intergenic sequence of a single organism that match one larger sequence of another organism; for example, two subsequences in rat that match one larger sequence in human. To minimize the noise in the datasets used for motif detection, such clusters are split into subclusters. Subclusters contain only a single subsequence of each ortholog (paralog; Figure 2-4: a more thorough description is given in Appendix A, §A.3.1). A subcluster is tagged by a profile, containing the IDs of the different subsequences composing this subcluster. The input dataset for motif detection (Figure 2-1) thus consists of the mammalian subsequences in a subcluster, together with the intergenic region of the corresponding *Fugu* ortholog.

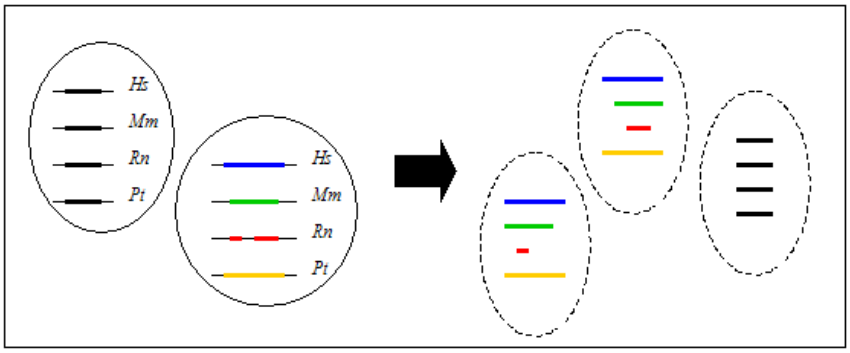


Figure 2-4. Schematic representation of subclusters, i.e., clusters of conserved orthologous sequences that contain one region in each ortholog. It is possible that a cluster contains different subsequences originating from the same intergenic region**.** This is the case for the (red) rat sequence in the left panel of the figure. Because these separated subsequences map to a different region of the intergenic sequence, they are not likely to contain the same regulatory motifs. Including them in the dataset for motif detection would increase the noise in the motif detection input set. In order to minimize the noise in datasets used for motif detection, such clusters were split into subclusters. A subcluster contains exactly one subsequence per ortholog (right panel). *Hs: Homo sapiens, Mm: Mus musculus, Rn: Rattus norvegicus, Pt: Pan troglotydes.*

2.4.2.2 Step II: Motif detection

To find motifs conserved in the preselected intergenic sequences of orthologous genes, we developed BlockSampler as an extension of MotifSampler (Thijs et al. 2002a). In contrast to the previous version of MotifSampler that could only handle a single background model, in BlockSampler each orthologous intergenic sequence in the input dataset is scored with its appropriate species-specific background model. Previous studies have shown that using the correct species-specific higher order background model improves the reliability of the results (Thijs et al. 2001; Marchal et al. 2003). In this study we used species-specific third-order background models.

The current implementation also allows selecting a user-defined core ortholog. This is the sequence of interest in which the motif should be present (in our case the sequence of heterogeneous length, i.e. the *Fugu* sequence). The idea behind this is that we are interested in motifs present in this core sequence that are supported by their presence in the preselected conserved orthologous regions. In this study, the most divergent *Fugu* orthologs were chosen as core sequences. The Gibbs sampling procedure searches for a common motif that has exactly one occurrence in the core sequence and no or one occurrence(s) in the remainder of the sequences. After short motif seeds are identified, these are extended using a simple protocol to find larger conserved blocks: if the consensus score over a 5-nt region adjacent to the current motif exceeds a given threshold, the motif is extended with one nucleotide (in that direction). The larger a conserved block, the higher the confidence in the motif.

BlockSampler was run 100 times for each input set (i.e. subcluster + *Fugu* ortholog) and corresponding random sets using default parameters - searching plus strand only (s=0), prior set to 0.2, initial motif length of 8 nt. Only the threshold of the consensus score (default 1.0) was augmented to 1.2, selecting stronger conserved blocks. This generated 100 conserved blocks for each input set. To avoid redundancy, blocks overlapping for more than 80% were merged. Concerning the benchmark datasets, consisting of only one distantly related ortholog namely *Fugu*, we then selected those blocks that were conserved among all vertebrates under study. When studying more diverse datasets, containing multiple distantly related species (with regard to mammals) we relaxed this requirement by allowing a block to be absent from one of the orthologs under study.

To account for the fact that short blocks are more likely to have a higher degree of conservation than long bocks, consensus scores (Thijs et al. 2002a) were compensated for their length. Blocks were then ranked according to this normalized consensus score (Cs_{ad}), calculated by the following formula

$$Cs_{ad} = \frac{L}{L+E} Cs \qquad [2.1]$$

where *L* is the length of the conserved block, *E* is an empirical factor (set to 5) and *Cs* the consensus score (also see Appendix A, §A.5.3).

To assess the relative contribution of individually data reduction and motif detection steps on the final result, we applied BlockSampler on the full-length benchmark datasets. We used the same parameter setting as described above, but because of the longer sequence length in the full datasets, we increase the number of runs (1000 runs for each benchmark

dataset). Blocks were selected as described above. The best scoring 10% of the remaining blocks were searched for known motifs.

2.4.3 Randomization

To set a threshold on the adapted consensus score of the blocks (blocks with a score above the threshold are considered relevant), we compared block scores of the genuine set with those of corresponding random sets. For each genuine dataset, 100 random sets were generated. A corresponding random set contains, besides the different homologous regions of the genuine subcluster under study, a random *Fugu* intergenic sequence. This additional random sequence was not orthologous with the mammalian sequences and thus is unlikely to contain the same motifs. In each random set, motifs were identified using the same procedure as described for the genuine set. For each random set the best scoring motif was selected, i.e. the block with the highest normalized consensus score. This resulted in a group of the best scoring 100 false positive motifs. These scores were approximately normally distributed. As a threshold, we choose the 90th percentile of the best scoring random motifs.

2.4.4 Motif validation

For each block we detected, a BLAT search against the human genome (assembly May 2004) was performed (Kent 2002; 2004a). This linked to the UCSC genome browser (Karolchik et al. 2003), where alignments between multiple vertebrate organisms are generated using MULTIZ (Blanchette et al. 2004). Subsequently, we checked in the UCR browser (Sandelin et al. 2004) whether ultra-conserved regions (UCRs) were identified in the intergenic regions under study.

To assess whether known transcription factor binding sites are located in the detected blocks, we compared the consensus sequence of each block with motifs described in the literature. In addition, we scanned the block consensus sequence with the TRANSFAC 6.0 public database of vertebrate transcription factor binding site profiles (Wingender et al. 2001). This scanning was performed using MotifLocator (Coessens et al. 2003; Aerts et al. 2003) with a 0^{th} order vertebrate background model. Hits with a score > 0.9 were regarded as potential binding sites. The binding sites are indicated by the TRANSFAC factor name (Wingender et al. 2001).

To calculate the statistical overrepresentation of homeodomain binding sites, 100 sequences were selected randomly from the *Fugu* genome and screened to make sure they differ from the genes under study. These

random sequences were screened with matrix models from homeodomain binding sites (obtained from TRANSFAC 8.2) using MotifLocator, as described above. We calculated the chi-square statistic with Yates correction of the 2 x 2 contingency table test for the set of homeodomain binding sites (Kato et al. 2004). Homeobox binding sites were significantly overrepresented in a certain block at the p-value of 10^{-8}.

2.4.5 Performance evaluation

To evaluate our newly developed procedure, we compared its performance to that of two algorithms often used for phylogenetic footprinting, namely a motif detection algorithm, MotifSampler (Thijs et al. 2002a) and the multiple alignment procedures MAVID (Bray et al. 2003; Bray and Pachter 2004) and TBA (Blanchette et al. 2004). These three algorithms were applied to the benchmark datasets and the resulting motifs (conserved in all organisms under study) were compared to those detected by the two-step procedure. We aligned the full-length initial datasets (Table 2-6 and Appendix A, §A.2: Table A-1) using the online MAVID version at with the default parameter setting (Bray and Pachter 2003).

Besides MAVID we used TBA as it has been shown to outperform MAVID (Blanchette et al. 2004). All the necessary tools were obtained from the Miller Lab website (2005c). To generate a multiple alignment using TBA we first pairwise aligned the initial datasets using BLASTZ. The following evolutionary tree was used: ((human chimp)(rat mouse) *Fugu*); the additional BLASTZ parameter file (latest version) was obtained from E. Margulies ftp site. The final multiple alignment was obtained by running the TBA executable.

We applied MotifSampler both on the reduced datasets (subcluster + complete *Fugu* intergenic sequence) and on the complete intergenic sequences (initial datasets). For the reduced sets we performed 100 MotifSampler runs while for the complete datasets MotifSampler was run 1000 times, each time using the standard parameter settings of the algorithm, i.e. the algorithm searches only for one motif (n=1) of 8 nt (w=8) on both strands (s=1), the prior probability of 1 motif copy (p) is 0.5. A third order vertebrate background model was used.

Chapter 3

Divergence of regulatory sequences in duplicated fish genes

3.1 Introduction

In this chapter we will develop a procedure for the identification of regulatory motifs in support of subfunctionalization and apply it to some control gene sets, i.e. groups of orthologous genes including paralogs that exhibit complementary expression patterns, i.e., subfunctionalization[1]. This chapter has been published as a book chapter in Genome Dynamics vol. 3 on 'Gene and Protein Evolution' (Van Hellemont et al. 2006).

When a gene gets duplicated, it awaits four possible fates. The most likely fate is pseudogenization or nonfunctionalization (Lynch and Force 2000; Lynch and Conery 2000; Maere et al. 2005). In rare cases, one of the two duplicates acquires a new function (neofunctionalization; (Taylor and Raes 2004)). Subfunctionalization, where both gene copies divide the gene's original functions, forms a third potential fate (Force et al. 1999). Furthermore, recent studies revealed that subfunctionalization is often accompanied by neofunctionalization, which has led to a new model of gene function evolution called sub-neofunctionalization (He and Zhang 2005). Finally, both copies can be retained, but, instead of diverging in function, they remain largely redundant and provide the organism with increased genetic robustness against harmful mutations (Gu 2003; Casneuf et al. 2006). In addition, retention and redundancy of genes, at least for certain functional classes, is predicted by the 'gene balance' hypothesis, which

[1] This research was performed in collaboration with research division 'Bioinformatics and Evolutionary Genomics' (Department of Plant Systems Biology, University of Ghent) under supervision of Prof. Yves Van de Peer.

states that retention of genes with strong dosage effects, such as for instance transcription factors, will be selected against if they are copied without their interacting partners (Maere et al. 2005; Blomme et al. 2006; Freeling and Thomas 2006).

In particular the subfunctionalization model (Force et al. 1999; Lynch and Force 2000) received much attention of late, since it can, at least partially, explain the large number of genes retained after duplication events, and their subsequent functional divergence (Moore and Purugganan 2005). The subfunctionalization model assumes that besides depending on its protein function, the functionality of a gene is also determined by its expression domain (where and when the gene is expressed). The specific expression domain of a gene results at least partially from its transcriptional regulation, which is, in turn, encoded by a specific combination of transcription factor binding sites (defined as a regulatory module) in the gene's promoter. Each transcription factor binding site (TFBS) in a module corresponds to a DNA consensus sequence or motif that is recognized by its cognate regulatory protein or transcription factor. Changes in these TFBS can therefore be an important antecedent for expression divergence and thus for sub- or neofunctionalization (Force et al. 1999; Prince and Pickett 2002; Gu et al. 2002; Papp et al. 2003; Zhang et al. 2004; Gu et al. 2005; Kafri et al. 2005; De Bodt et al. 2006; Casneuf et al. 2006). However, the number of studies that show how expression divergence between paralogs is reflected by differences between regulatory elements of paralogous gene pairs is still limited (Chang et al. 2006; Jimenez-Delgado et al. 2006).

As a result of a genome wide fish-specific duplication event that occurred some 350 mya (Vandepoele et al. 2004; Jaillon et al. 2004; Christoffels et al. 2004) and of more recent duplication events (Blomme et al. 2006), ray-finned fish such as *Tetraodon* and zebrafish contain a large number of duplicated genes (Postlethwait et al. 1998; Wittbrodt et al. 1998; Van de Peer Y. et al. 2003; Meyer and Van de Peer Y. 2005), of which several have already been shown to have undergone subfunctionalization (Postlethwait et al. 1998; Van de Peer et al. 2001; Altschmied et al. 2002; Winkler et al. 2003; Postlethwait et al. 2004; Volff 2005; Bollig et al. 2006). In this study, we further investigate to what extent '*in silico*' analyses support expression divergence of genes through identified changes in regulatory sequences. To this end, we explicitly searched for motifs that have been preserved over 450 mya of vertebrate evolution (from mammals to ray-finned fish), but have been differentially retained in either one of two duplicates in zebrafish or *Tetraodon* and that thus might explain the experimentally observed expression divergence.

3.2 Results

The goal of our study was to see whether divergent expression of duplicated genes is reflected in any detectable way by a different composition of their regulatory sequences, in particular by the presence or absence of specific motifs. First, this requires identifying interesting case studies, i.e., gene families that contain members of fish specific duplication events. Second, we need to compile the potential regulatory motifs present in the intergenic sequences of these gene families and to identify which of the motifs have been differentially retained in one of the fish paralogs. However, since the regulatory motifs present in fish genomes are still largely unknown, the list of potential motifs was compiled based on comparative *de novo* motif detection methods, better known as phylogenetic footprinting. Phylogenetic footprinting assumes that biologically relevant sequences, such as regulatory motifs, evolve slower than their surrounding non functional intergenic sequences. By using cross-species conservation, short stretches of DNA that are conserved over certain phylogenetic distances are identified as potential motifs. The greater the phylogenetic distance over which the motif is conserved and the more orthologs in which the motif can be detected, the more confidence can be put in this prediction.

However, as we are specifically searching for conserved motifs that are differentially lost between paralogs, we had to rely on a phylogenetic footprinting methodology that is able to align strongly evolved sequences of which some do not contain the motifs (Van Hellemont et al. 2005).

3.2.1 Identifying gene sets containing duplicate fish genes

Our analysis was performed on a selection of gene families that contained paralogs either originating from a duplication event before the divergence of zebrafish and *Tetraodon* (further referred to as the ancient fish specific duplication, FSD) or from a more recent duplication specific to either zebrafish or *Tetraodon* (that occurred after divergence of both species (Blomme et al. 2006)) (for a complete list of these gene families see Table 3-1). Here follows a short description of their functions: *bmp2* encodes the bone morphogenetic protein 2 precursor, which functions in bone formation. The ephrin-A1 precursor, encoded by *efna1*, belongs to the ephrin family of receptor tyrosine kinases ligands. The gene product of *en2* is the homeobox protein engrailed-2. This developmental protein is expressed during vertebrate embryogenesis. Glycine receptors, encoded by *glyR* genes, are ligand-gated ion-channels that mediate inhibitory synaptic transmission in the brainstem and the spinal cord of vertebrates. *kcnip1* encodes the

potassium channel-interacting protein 1, KChIP1. The homeobox protein Msx2 has a role in vertebrate morphogenesis. *ntng1* and *ntng2* encode respectively Netrin-G1 precursor and Netrin-G2 precursor, which are both involved in the outgrowth of axons and dendrites. Pax6 functions in the central nervous system and in the developing eye. *ssh* encodes the Sonic hedgehog protein precursor, which is essential for a variety of patterning events during vertebrate development. Finally, the developmental homeobox protein Six4 plays a role in anatomical structure morphogenesis.

For some of these fish-specific paralogs, subfunctionalization was supported by literature (e.g., *bmp2* (Martinez-Barbera et al. 1997; Laforest et al. 1998), *glyRα* (Imboden et al. 2001), *msx2* (Ekker et al. 1997), *pax6* (Nornes et al. 1998; Force et al. 2004) and *ssh* (Laforest et al. 1998) (Table 3-1: marked with an asterisk)).

Table 3-1. Description of the datasets. Dataset: indicates the name of the gene family (derived from the human ortholog in the dataset). For gene sets indicated with an asterisk, experimental evidence supporting expression divergence between the fish paralogs exists. Newick tree: for each dataset the phylogenetic relations are given in Newick format. Dr: *Danio rerio*; Gg: *Gallus gallus*, Hs: *Homo sapiens*, Tn: *Tetraodon nigroviridis,* Xt: *Xenopus tropicalis*. Ensembl Gene IDs: lists the genes present in each dataset by their Ensembl gene id. Experimental evidence: indicates the type of experimental evidence that supports expression divergence.

Dataset		Newick tree	Ensembl Gene IDs	Experimental evidence
bmp2 *	*	((Xt, (Hs, Gg)), (Dr1, (Dr2, Tn)));	(Dr1) ENSDARG00000013409, (Dr2) ENSDARG00000041430, (Gg) ENSGALG00000008830, (Hs) ENSG00000125845, (Tn) GSTENG00020275001, (Xt) ENSXETG00000005519	RT-PCR + *in situ* hybridization (Martinez-Barbera et al. 1997)
efna1		(Hs, ((Dr1, Tn1), (Dr2, Tn2)));	(Dr1) ENSDARG00000030326, (Dr2) ENSDARG00000018787, (Hs) ENSG00000169242, (Tn1) GSTENG00032578001, (Tn2) GSTENG00033951001	-
en2		((Xt, Hs), (Dr1, (Dr2, Tn));	(Dr1) ENSDARG00000026599, (Dr2) ENSDARG00000038868, (Hs) ENSG00000164778, (Tn) GSTENG00023985001, (Xt) ENSXETG00000013496,	-
glyRα1 *	*	(Gg, ((Dr, Tn1), Tn2));	(Dr) ENSDARG00000006865, (Gg) ENSGALG00000004936, (Tn1) GSTENG00029286001, (Tn2) GSTENG00022245001	*In situ* hybridization (Imboden et al. 2001)
glyRα1-related		((Xt, Gg), (Dr1, (Dr2, Tn)));	(Dr1) ENSDARG00000012019, (Dr2) ENSDARG00000011066, (Gg) ENSGALG00000004134, (Tn) GSTENG00024269001, (Xt) ENSXETG00000001966	-

kcnip1		((Xt, (Gg, Hs)), ((Dr1, Tn1), (Dr2, Tn2)));	(Dr1) ENSDARG00000034808, (Dr2) ENSDARG00000022109, (Gg) ENSGALG00000002132, (Hs) ENSG00000182132, (Tn1) GSTENG00020358001, (Tn2) GSTENG00024581001, (Xt) ENSXETG00000018293	-
msx2 *	*	((Xt, (Gg, Hs)), (Dr1, Dr2));	(Dr1) ENSDARG00000009936, (Dr2) ENSDARG00000006982, (Gg) ENSGALG00000002947, (Hs) ENSG00000120149, (Xt) ENSXETG00000009168	*In situ* hybridization (Ekker et al. 1997)
ntng1		((Gg, Hs), (Tn1, (Dr, Tn2));	(Dr) ENSDARG00000014973, (Gg) ENSGALG00000001896, (Hs) ENSG00000162631, (Tn1) GSTENG00027711001, (Tn2) GSTENG00035109001	-
ntng2		((Hs, Gg), ((Dr, Tn1), Tn2));	(Dr) ENSDARG00000036938, (Gg) ENSGALG00000003677, (Hs) ENSG00000196358, (Tn1) GSTENG00004089001, (Tn2) GSTENG00014392001	-
pax6 *	*	(((Gg, Hs), Xt), ((Dr1, Dr2), Tn));	(Dr1) ENSDARG00000045045, (Dr2) ENSDARG00000045936, (Gg) ENSGALG00000012123, (Hs) ENSG00000007372, (Tn) GSTENG00025814001, (Xt) ENSXETG00000008175	*In situ* hybridization + transient transfection assays + western blot analysis (Nornes et al. 1998)
shh *	*	((Hs, Gg), (Dr1, (Dr2, Tn));	(Dr1) ENSDARG00000038867, (Dr2) ENSDARG00000039710, (Gg) ENSGALG00000006379, (Hs) ENSG00000164690, (Tn) GSTENG00023991001	*In situ* hybridization (Laforest et al. 1998)
six4		((Xt, Hs), ((Dr1, Tn), Dr2));	(Dr1) ENSDARG00000031983, (Dr2) ENSDARG00000004695, (Hs) ENSG00000100625, (Tn) GSTENG00032223001, (Xt) ENSXETG00000016941	-

For gene sets *bmp2*, *efna1*, *en2*, *glyRα1*, *glyRα1*-related, *kcnip1*, *ntng1*, *ntng2*, *pax6*, *shh* and *six4,* the topology of the corresponding phylogenetic trees indicates that the paralogs resulted from the ancient FSD event which took place before the divergence of zebrafish and *Tetraodon* (about 150 mya (Meyer and Van de Peer Y. 2005)). For the *pax6* gene family, the two zebrafish copies are the result of a more recent zebrafish specific duplication event (Table 3-1). Concerning the *msx2* gene family, the topology did not allow us to conclude whether the *msx2* zebrafish copies resulted from the ancient FSD or whether they were the result of a more recent duplication event in zebrafish.

3.2.2 Determining the overall homology between paralogous intergenic regions

In order to determine their overall conservation, intergenic paralogous regions in fish were aligned using Smith-Waterman (Smith and Waterman 1981). Results are shown in Table 3-2. The conservation level of paralogous intergenic regions resulting from the ancient FSD (Blomme et al. 2006) (*Tetraodon* 40.4% and *Danio rerio*: 43.6%) was comparable to that of unrelated sequences, which was estimated 40.4% and 43% for *Tetraodon* and *Danio rerio* respectively (see Methodology, §3.4.2). This analysis also indicates that in these ancient duplicates, except for the conserved regulatory motifs, no sequence conservation is to be expected.

Table 3-2. Intergenic homology between fish duplicates. Intergenic regions of fish paralogs present in the gene sets under study were pairwise aligned using Smith-Waterman (1981). Dataset: indicates the name of the gene family (derived from the human ortholog in the dataset) from which the fish paralogs were compared. For gene sets indicated with an asterisk, experimental evidence supporting expression divergence between the fish paralogs exists. Compared duplicates: the fish genes for which the intergenic sequences were aligned; for the corresponding Ensembl gene ids we refer to Table 3-1. Dr: *Danio rerio*; Gg: *Gallus gallus*, Hs: *Homo sapiens*, Tn: *Tetraodon nigroviridis,* Xt: *Xenopus tropicalis.* Type: the type of duplication event the duplicates are the result of (based on the phylogenetic trees); FSD: ancient fish specific duplication; ZS: zebrafish specific duplication. Percent identity: the similarity between the intergenic sequences of the aligned duplicates.

Dataset		Compared duplicates	Type	Percent Identity
bmp2	*	Dr1 –Dr2	FSD	44.4%
glyRα1	*	Tn1 –Tn2	FSD	41.4%
msx2	*	Dr1 – Dr2	FSD/ZS	44.5%
pax6	*	Dr1 – Dr2	ZS	39.2%
shh	*	Dr1 – Dr2	FSD	43.6%
efna1		Dr1 – Dr2	FSD	43.9%
		Tn1 – Tn2	FSD	32.6%
en2		Dr1 – Dr2	FSD	45.9%
glyRα1-related		Dr1 – Dr2	FSD	42.3%
kcnip1		Dr1 – Dr2	FSD	43.4%
		Tn1 – Tn2	FSD	41.9%
ntng1		Tn1 – Tn2	FSD	42.3%
ntng2		Tn1 – Tn2	FSD	43.9%
six4		Dr1 – Dr2	FSD	41.7%

3.2.3 Identification of motifs supporting subfunctionalization

To search for differences in motif composition between duplicates, we first compiled all potential motifs conserved within the intergenic regions of genes belonging to the same gene family (see Methodology, §3.4.3). To this end we used BlockSampler, a procedure based on Gibbs sampling that searches for statistically overrepresented motifs. In the presence of an appropriate background model, the procedure is known to be quite robust against noise, i.e., sequences that do not contain the motif (Thijs et al. 2001; Thijs et al. 2002; Marchal et al. 2003). In the context of subfunctionalization, this property is essential as it allows finding motifs that are not conserved in all branches of the phylogenetic tree. For each set, applying BlockSampler resulted in a list of conserved motifs. To detect motifs supporting subfunctionalization, from this list we selected those motifs that were significantly conserved but missing in at least one of the fish paralogs, either *Tetraodon* or zebrafish (i.e., the species for which multiple paralogs are present in the gene set under study). Especially for the ancient duplications for which the overall similarity in intergenic sequences between the fish paralogs is quite low, many differences are expected to be found in their promoter regions, most of which probably do not correspond to biologically relevant subfunctionalized motifs. Therefore, in order to select the most relevant predictions we considered only those motifs that were also conserved in phylogenetic lineages other than fish and thus were preserved over 450 mya of vertebrate evolution (see Methodology, §3.4.4 for the exact criteria).

To test to what extent the choice of the threshold on the motif scores (defined as the xth percentile of the random scores) determined the total number of motifs retrieved and thus the number of gene sets for which we detected (a) motif(s) indicative for subfunctionalization, we repeated the analysis for multiple threshold levels (ranging from the 99.5^{th} to the 50^{th} percentile of the random score distribution). The results for gene sets containing homologs from multiple non-fish species are summarized in Table 3-3, considering the 99.5^{th}, 99^{th}, 95^{th} and 90^{th} percentile of the random distribution as threshold on the motif scores. As expected, lowering the threshold of our search allows detecting more motifs indicative for subfunctionalization. However, as the stringency of the search becomes lower, the motifs taken into account become gradually shorter and presumably less reliable.

Table 3-3. Motifs indicative for subfunctionalization with a Cs_{ad} score exceeding the 90th percentile of the random score distribution. Dataset*:* the gene set in which the motifs were detected. Gene sets indicated with an asterisk contain fish paralogs for which subfunctionalization has been shown in literature. #: the number of motifs detected in the gene set that support subfunctionalization given this threshold. L: the length of the motif indicative for subfunctionalization. PI: indicates the percentile of the random distribution to which the score of the motif belongs. Conservation profile: indicates in which homologs of the gene family the motif was also present (referring to Table 3-1). Dr: *Danio rerio*; Gg: *Gallus gallus*, Hs: *Homo sapiens*, Tn: *Tetraodon nigroviridis,* Xt: *Xenopus tropicalis.* Subf sp.: indicates in which fish-species the motif was lost. Duplication type: indicates from which duplication event the paralogs originated for which a motif was found that supports subfunctionalization. FSD: ancient fish specific duplication event; TS: *Tetraodon* specific duplication event; ZS: zebrafish specific duplication event. Motif name: the name to unambiguously indicate a specific motif.

Dataset		#	L	PI	Conservation profile	Subf sp.	Duplication type	Motif name
bmp2	*	1	29	99.5	Dr2_Gg_Hs_Tn	Dr	FSD	bmp2_1
pax6	*	4	70	90	Dr2_Gg Hs_Tn_Xt	Dr	ZS	pax6_1
			38	90	Dr2_Gg Hs_Tn_Xt	Dr	ZS	pax6_2
			71	90	Dr2_Gg Hs_Tn_Xt	Dr	ZS	pax6_3
			50	90	Dr2_Gg Hs_Tn_Xt	Dr	ZS	pax6_4
shh	*	2	15	95	Dr2_Gg_Hs_Tn	Dr	FSD	shh_1
			16	99	Dr2_Gg_Hs_Tn	Dr	FSD	shh_2
kcnip1		1	13	95	Dr1_Dr2_Gg_Hs_Tn1_Xt	Tn	FSD	Kcnip1_1
Total		**8**						

When using a quite conservative threshold (motif scores exceeding the 90th percentile of the random distribution), in three (*bmp2*, *pax6* and *shh*) out of the five datasets for which expression divergence was experimentally demonstrated, we could find at least one motif indicative for subfunctionalization. Besides in these experimentally supported datasets, we also found motif-based indications for subfunctionalization in *kcnip1* and *efna1* (although only when using a relaxed threshold in the motif scores, Table 3-4).

Table 3-4. Motifs indicative for subfunctionalization for the gene sets for which a relaxed selection criterium was used. For legend see Table 3-3.

Dataset	#	L	PI	Conservation profile	Subf sp.	Duplication type	Motif name
efna1	2	12	75	Dr1_Dr2_Hs_Tn1	Tn	TS	efna1_1
		8	55	Dr1_Dr2_Hs_Tn1	Tn	TS	efna1_2
Total	**2**						

3.2.4 Detailed description of the datasets with subfunctionalized motifs

Figures 3-1 to 3-5 display the results for the datasets *bmp2*, *pax6*, *shh, kcnip1* and *efna1*. Significantly overrepresented motifs are mapped. An arrow indicates motifs that might be supportive of subfunctionalization. Below we give a more detailed description of these results.

In vertebrates, bone morphogenetic proteins **(Bmps)** play a crucial role in establishing the early body plan and in organogenesis (Hogan 1996). Martinez-Barbera et al. (1997) studied the expression pattern of zebrafish *bmp2* paralogs, *bmp2a* and *bmp2b*. *In situ* hybridization showed a divergent expression profile for both paralogs in the gastrulating embryo and in the pectoral fin bud. In this study, the *bmp2* gene family consist of two zebrafish genes (Table 3-1, Figure 3-1) that correspond to the genes studied by Martinez-Barbera et al. (1997). The motif indicated in Figure 3-1 (Table 3-3) that has been retained in one zebrafish copy (Dr2, ENSDARG00000041430) but that was lost in the other (Dr1, ENSDARG00000013409), could possibly explain this observed divergence.

Pax6 plays an important role in the central nervous system and in the developing eye of both vertebrates and invertebrates. According to our analysis, the *pax6* gene family contains two zebrafish paralogs, which (given the position of the *Tetraodon* homolog in the tree topology) originated from a zebrafish specific duplication (Figure 3-2). The presence of two zebrafish paralogs is consistent with the observations of Nornes et al. (1998). They observed that both zebrafish copies have unique expression domains that sum up to the total expression domain for the single *pax6* copy present in birds and mammals (Lynch and Force 2000). Figure 3-2 displays the conserved motifs identified in the promoter region of the *pax6* homologs. Motifs that might be indicative for subfunctionalization are indicated by arrows: we identified four motifs conserved in human, chicken, frog,

Tetraodon and in one zebrafish paralog (Dr2, ENSDARG00000045936) (Table 3-3). The complete absence of all of these motifs in the zebrafish paralog (Dr1, ENSDARG00000045045) can also be interpreted as an indication for nonfunctionalization of this paralog. However, because experimental evidence about the expression of both zebrafish paralogs exists (Nornes et al. 1998), subfunctionalization seems the more likely fate. The order in which the motifs occur in the intergenic regions seems to be perfectly conserved in the non-mammalian sequences where they are concatenated into a large conserved region of circa 250 nt (Figure 3-2). In the human ortholog on the contrary, the order and spacing of these motifs seems to be altered. In order to get an idea of the binding sites localized in the four motifs reported here, we screened the human motif instance with the TRANSFAC database of transcription factor binding sites. As is summarized in Table 3-5, different potential binding sites are present in the *pax6* motifs. For instance, pax6_3 contains an AP-2α and an AP-2rep binding site. This is plausible, since both Pax6 as AP-2 α function in eye development (West-Mays et al. 1999). Moreover, both transcription factors are known to interact in coordinating corneal epithelial repair (Sivak et al. 2004).

Vertebrate **hedgehog** genes are involved in many developmental processes (Ingham and McMahon 2001). As Laforest et al. (1998) showed that zebrafish hedgehog paralogs exhibit expression patterns that suggest subfunctionalization, we choose to study the *sonic hedgehog* or *shh* gene family (Laforest et al. 1998) in more detail. Figure 3-3 illustrates two significant motifs (see arrows) that possibly support subfunctionalization (see also Table 3-2). These motifs, indicated in red and green, have both been conserved in human, chicken, *Tetraodon* and Dr2 (ENSDARG00000039710) but were lost in Dr1 (ENSDARG00000038867). The order and spacing between these two motifs seems also to have been retained during evolution. Besides these, Figure 3-3 also displays some additional interesting motifs pointing towards subfunctionalization (for instance, the dark purple and dark yellow motifs). These were not initially retained as "significant motifs" under our strong selection criteria, because they were either too short or not conserved in multiple non-fish species.

kcnip1 encodes the potassium channel-interacting protein (Shibata et al. 2003). In this study we identified a frog, chicken and human homolog, two zebrafish and two *Tetraodon* paralogs. The tree topology (Figure 3-4) indicates that the four fish genes are the result of an ancient FSD. As is shown in Figure 3-4, we identified one motif that is in support of a possible divergent expression profile (Table 3-3). This motif, indicated in red, is retained in frog, human, both zebrafish paralogs, and one *Tetraodon* paralog (Tn2, GSTENG00024581001). Two other interesting motifs (the green and light blue motifs) were detected in this dataset: both these motifs seem to be present in the human sequence but are divergently retained over the fish

paralogs. The smaller blue motif is retained in Dr1 and Tn1, while the green motif is retained in Dr2 and Tn2. From the tree topology it seems that the combined motif, still present in the human sequence might have been subfunctionalized after an early fish duplication that took place before the speciation between *Tetraodon* and zebrafish. The motifs seem to have a classical pattern of subfunctionalization. Note, however, that we did not primarily retain them, as they do not meet our selection criteria (the motifs are conserved in one non-fish homolog only).

Also in ***efna1***, which encodes an ephrin-A1 precursor, we found two motifs indicative of expression divergence between paralog Tn1 (GSTENG00032578001) and paralog Tn2 (GSTENG00033951001) (Figure 3-5). These two motifs, respectively 12 and 8 bp long, and conserved in the intergenic sequences of the human homologue were also present in both zebrafish paralogs, but only in one of the two *Tetraodon* paralogs (Tn1) (see Table 3-4).

Table 3-5. The potential transcription factor binding sites located in the detected motifs indicative for subfunctionalization (Table 3-3 and). Motif name*:* the name to unambiguously indicate a certain motif detected (with a Cs_{ad} score exceeding 90^{th} percentile of the random score distribution for *bmp2*, *pax6*, *ssh* and *kcnip1* and a Cs_{ad} score exceeding 50^{th} percentile for *efna1*). These names correspond to the ones in Table 3-3 and Table 3-4. Consensus and possible binding sites: the sequence of the motif in the intergenic region of the human homolog (Table 3-1) is given followed by the possible binding sites situated in this motif (TRANSFAC name, TRANSFAC ID, consensus sequence, positions, strand and score). Remark: For the shortest motifs MotifLocator could not be used to screen for potential binding sites. Therefore only the motif instance in human is given.

Motif Name	Consensus and possible binding sites
bmp2_1	**TTGTTTTGTTTTGTTTTTT**
	SRY, M00148: AAACWAM: 5-11- (1.0); 10-16 - (1.0)
pax6_1	**GGCTCGAGGGCCAGGTTGAGGGTACTCATCGAGCCTCGAACTCCTCCTAAAAATGATTCCTGCCAAAAGC**
	Cap, M00253, NCANHNNN: 49-56 - (0.963)
	CdxA. M00101, AWTWMTR: 46-52 - 0.904)
	Hnf4, M00967, AARGTCCAN: 6-14 + (0.931)
	Etf, M00695, GVGGMGG: 43-49 - (0.906)
	Lyf-1, M00141, TTTGGGAGR, :59-67 - (0.931); 43-51 - (0.950)
	NF1, M00193, NNTTGGCNNNNNNCCNNN: 51-68 - (0.919)
pax6_2	**ACCACTGTCACTTTCAAATTGGAGAGCCAGATGGAAGC**
	E2a, M00804, GGCGSG: 21-34 - (0.907)
	Irf, M00772, BNCRSTTTCANTTYY: 1-15 + (0.946)
	Tal1, M00993, TCCAKCTGNY: 26-35 - (1.0)
pax6_3	**TGGTAAGGTCTAGGCCCAGACTAGAGTGGCCAGTGGGAGGTGGGCGCTCCTAGGCCTTAACACAGGATGCC**
	AP-2α, M00469, GCCNNNRGS: 29-37 + (0.917)
	AP-2rep, M00468, CAGTGGG: 31-37 + (1.0)
	Cap, M00253, NCANHNNN: 16-23 + (0.906); 63-70 - (0.925); 36-43 - (0.904)
	CCAAT box, M00254, NNNRRCCAATSA: 25-36 + (0.933)
	C/EBP, M00159, NNTKTGGWNANNN: 54-66 - (0.906)
	CHCH, M00986, CGGGNN:34-39 + (0.932)
	Etf, M00695, GVGGMGG: 34-40 + (0.911)
	Ets, M00971, ACTTCCTS: 63-70 - (0.925)
	Pea3, M00655, ACWTCCK: 64-70 - (0.933)
pax6_4	**ATTTTCCTGTTTTCCTCCTCTAAGTCACAAAGTCAACAGTTAATTCAAAG**
	AP-1, M00172, RSTGACTNMNW: 19-29 - (0.942)
	AP-1, M00517, NNNTGAGTCAKCN: 18-30 - (0.905)
	AP-1, M00924, TGACTCANNSKN:16-27 - (0.901)
	AP-1, M00926, TGAGTCAN:21-28 + (0.904);
	Fox, M00809, KATTGTTTRTTTW: 29-41- (0.953)
	Hnf3α, M00724, TRTTTGYTYWN: 28-38 - (0.932)
	Hnf3β, M00131, KGNANTRTTTRYTTW: 29-43 - (0.908)
	Pou1f1, M00744, ATGAATAAWT: 39-48 - (0.915)
	Sf1, M00727, TGRCCTTG: 28-35 - (0.918)
	Stat1, M00496, NNTTTCCN:1-8 + (0.948); 9-16 + (0.9704)
	Stat6, M00500, NNYTTCCY : 9-16 + (0.915)
shh_1	**GCTCTCCAGGCTTGC**
shh_2	**TCAGATGCGCCCCTGG**
kcnip1_1	**TGTGTATCTGTGT**
efna1_1	**ACGCAGACACACA**
efna1_2	**ATGTTTATT**

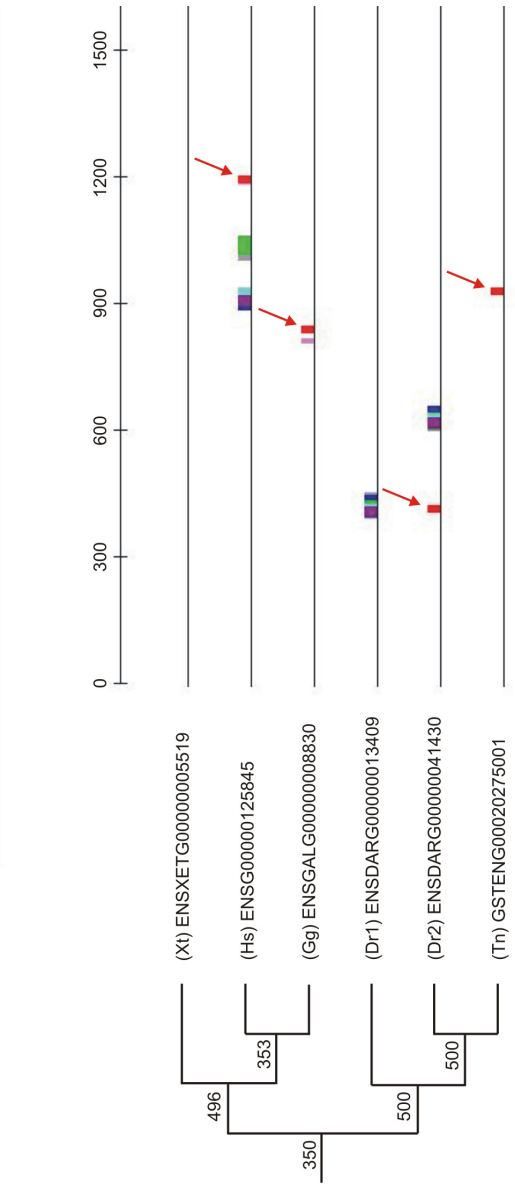


Figure 3-1. Graphical display of the motifs found in the upstream regions of *bmp2*. Graphical display of all motifs with a score exceeding the 90th percentile of the random score distribution. The motifs that are in support of subfunctionalization are indicated by an arrow. The phylogenetic tree (branch lengths not drawn to scale) illustrates the evolutionary relationships between the homologs; these are indicated as defined in Table 1. The bootstrap values are indicated on the branches of the phylogenetic tree. Abbreviations used: Xt: *Xenopus tropicalis*, Hs: *Homo sapiens*, Gg: *Gallus gallus*, Dr: *Danio rerio*, Tn: *Tetraodon nigroviridis.*

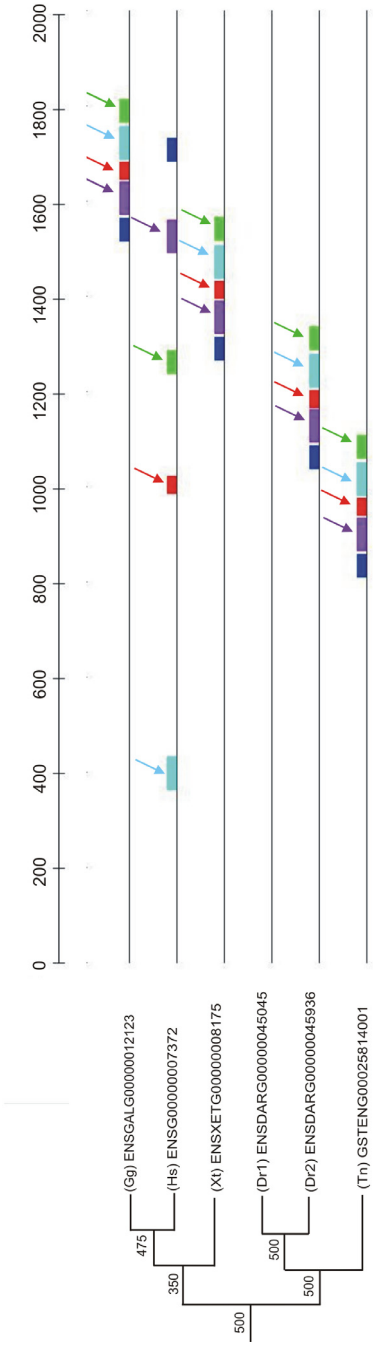


Figure 3-2. Graphical display of the motifs found in the upstream regions of *pax6*. Interpretation is as in Figure 3-1.

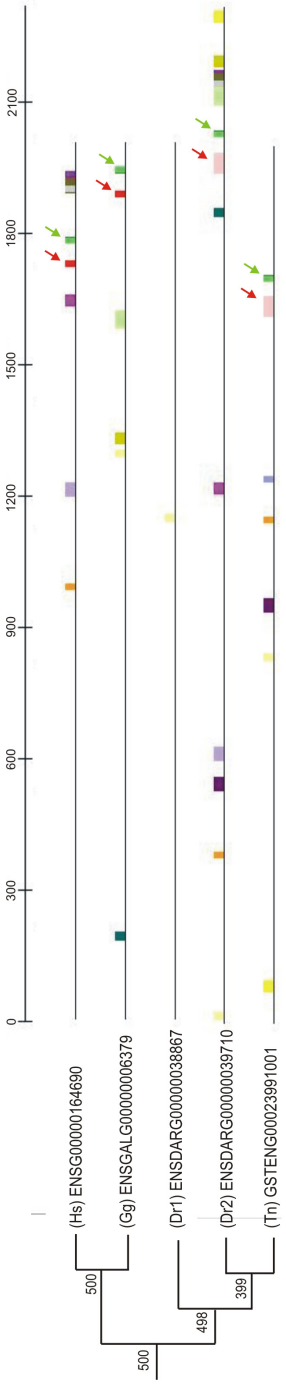


Figure 3-3. Graphical display of the motifs found in the upstream regions of *shh*. Interpretation is as in Figure 3-1.

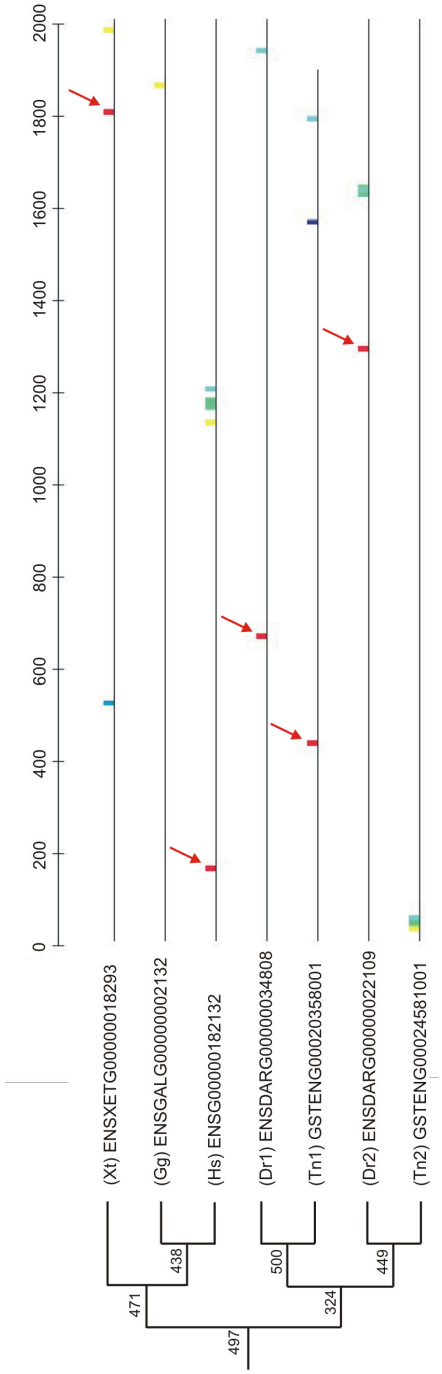


Figure 3-4. Graphical display of the motifs found in the upstream regions of *kcnip1*. Interpretation is as in Figure 3-1.

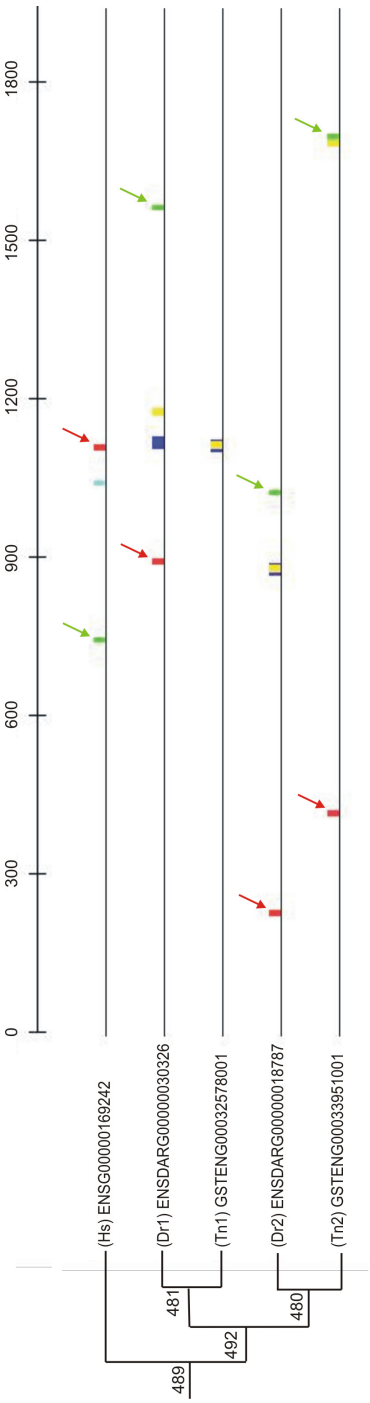


Figure 3-5. Graphical display of the motifs found in the upstream regions of *efna1*. Interpretation is as in Figure 3-1.

3.3 Discussion

In this study, we found indications that expression divergence between paralogs in zebrafish and/or *Tetraodon* is reflected by differences in regulatory motifs. We investigated five gene families for which experimental evidence supported subfunctionalization. For three of these proof-of-concept gene families, we identified at least one motif that was differentially lost after a fish-specific duplication event and that seemed to be in accordance with the experimentally observed expression divergence. Besides in the 'proof of concept' datasets, we found differential alterations in regulatory motifs between the fish paralogs that point towards potential subfunctionalization in two other gene families (*efna1* and *kcnip1)*.

In order to assess which potential transcription factors bind to the conserved motifs, we screened them with the TRANSFAC database of TFBS. Several potential TFBS seemed to be present in the conserved motifs but to our knowledge, for the majority of these TFBS no clear link with the genes containing the conserved motifs was found in literature.

The sequence dependent indications for subfunctionalization identified in this study of course largely depend on the reliability of our *in silico* predicted regulatory motifs. To select confident predictions, we used strict selection criteria and considered only those motifs that were conserved over at least 450 mya of vertebrate evolution. On the other hand, by using these conservative selection criteria we probably discard many functional motifs. Indeed, motifs that are too short, too degenerated, or very lineage specific will remain undetected. As a result, we most likely underestimate the number of motifs indicative for subfunctionalization. This might explain why we only find in a subset of the 'proof of concept' datasets sequence based indications for subfunctionalization. Moreover, according to its strict definition, subfunctionalization implies that both paralogs divide the gene's original function over both gene copies. When relating this to expression divergence and subsequent changes in regulatory motifs, one expects to find two ancestral motifs still present in an outgroup species to be divided between the two paralogous intergenic regions. We could not detect any example of this idealized situation of subfunctionalization due to our conservative approach; however, when using more relaxed criteria our method identified such an example.

Despite our conservative strategy, in nearly half of the tested datasets clear sequence based indications for potential expression divergence were present, indicating that subfunctionalization is probably more general than is assumed at this point (see also Casneuf et al. 2006).

3.4 Methodology

3.4.1 Identification of suitable datasets

In this study we focused on duplicated fish genes for which there was (some) experimental evidence that supported subfunctionalization, such as for *bmp2* (Martinez-Barbera et al. 1997), *glyRα* (Imboden et al. 2001), *msx2* (Ekker et al. 1997), *pax6* (Nornes et al. 1998; Force et al. 2004), and *shh* (Laforest et al. 1998). In addition, several other duplicated fish genes were included, namely *efna1*, *en2* (Joyner and Martin 1987; Force et al. 1999), *kcnip1*, *ntng1*, *ntng2* and *six4,* and a *glyRα*-related gene family. Phylogenetic trees were constructed based on the predicted protein sequences (Ensembl release 37 at http://www.ensembl.org) from human, mouse, *Tetraodon nigroviridis*, zebrafish (*Danio rerio*), rat (*Rattus norvegicus),* chicken (*Gallus gallus*), and frog (*Xenopus tropicalis*). If splice variants were reported, the longest transcript was used.

To delineate vertebrate gene families, a similarity search was performed (BLASTP, (Altschul et al. 1997); E-value cut-off E-10) with all proteins from the organisms listed above, plus those of *Ciona intestinalis* (JGI, http://genome.jgi-psf.org), version 1 and *Drosophila melanogaster* (Ensembl), version 3, which were added as outgroup species. Blast hits between vertebrate sequences with a score better than the best score between a vertebrate sequence and an outgroup sequence (*Drosophila* or *Ciona)* were retained and considered members of the same gene family. The *Drosophila* or *Ciona* sequence was used to root the phylogenetic tree (see further). For each gene family, a multiple sequence alignment was created with T-Coffee 1.37 using default parameters (Notredame et al. 2000). Alignment columns containing gaps were removed when a gap was present in more than 10% of the sequences. To reduce the chance of including misaligned amino acids, all positions in the alignment left or right from the gap were also removed until a column in the sequence alignment was found where the residues were conserved in all genes included in our analyses. This was determined as follows: for every pair of residues in the column, the BLOSUM62 value was retrieved. If at least half of the pairs had a BLOSUM62 value ≥ 0, the column was considered as conserved. Neighbor joining trees (with 500 bootstrap replicates) were constructed with PHYLIP 3.5 (Felsenstein 1989) using both nucleic and amino acid sequence alignments and simple poisson-corrected substitution models.

For all datasets, the phylogenetic tree showed genes duplicated in at least one of the fish species. These duplicates, and their homologs in human, chicken and frog, were selected for further analysis (Table 3-1). Intergenic regions of these homologs were retrieved using the Ensembl mart database

release 37. Intergenic regions were defined as the region upstream of the transcription start (as defined by Ensembl), limited to 2 kb and including the 5'UTR.

3.4.2 Pairwise alignment of paralogous intergenic sequences

Pairwise alignments of paralogous intergenic regions were obtained with Smith-Waterman using the default parameters (gap open penalty 10 and gap extension penalty 0.5) (Smith and Waterman 1981). The expected percentage identity between two unrelated intergenic sequences of the same organism was estimated by averaging the scores obtained by aligning each intergenic sequence against all other intergenic sequences of the same organism but not belonging to the same protein family. These calculations were used in §3.2.2.

3.4.3 Search for regulatory motifs conserved in each of the gene sets

3.4.3.1 Motif detection

For each gene set (i.e., all genes belonging to the same gene family), intergenic sequences were subjected to BlockSampler (Van Hellemont et al. 2005). BlockSampler was run using default parameters - searching plus strand only (s=0) and searching for one motif per run, prior set to 0.2, initial motif length of 8 nt, and a threshold on consensus score of 1.0. BlockSampler requires the definition of a root sequence, i.e. only conserved motifs, which are also present in the root, will be retained. Because for our application the biological meaning of a root was less clear, each sequence of the gene set was chosen once as root. Per root sequence BlockSampler was run 100 times implying that the total number of runs and retrieved motifs for a gene set equalled 100 times the 'number of sequences in the gene set'. The motifs with a consensus score above 1 were selected and motifs overlapping for more than 80% were merged to avoid redundancy.

In order to account for the fact that short motifs are more likely to have a higher degree of conservation than long motifs, the consensus score of each detected block was normalized for the length of the motif using formula [2.1].

3.4.3.2 Assessing the statistical significance of detected motifs

For each gene set, 30 random sets were compiled. These random sets have a composition similar to the genuine gene set in sequence number and origin (species), but in contrast to the genuine gene set sequences were selected randomly and as a result do not share any homology relation. For each random set, we performed the same analysis as for the genuine gene sets: BlockSampler was applied to identify conserved motifs. Per random set, the number of runs equalled 100 times the number of sequences in the random set (of which each one served once as root). After normalizing the scores, from the 100 runs of a single root the best scoring motif (highest Cs_{ad}) was selected. This resulted, for each genuine gene set, in a number of random motifs equalling 30 (i.e., the number of random sets) times the 'number of sequences in this random set (i.e., number of root sequences)'. The scores of these motifs were used to estimate a random motif score distribution. To identify significant motifs in the genuine dataset we chose the Cs_{ad} of the x^{th} percentile of this random distribution as a threshold. As a result, motifs in the genuine dataset with a Cs_{ad} higher than the chosen threshold were considered statistically significant.

3.4.4 Identifying motifs supporting subfunctionalization

Motifs that potentially support the subfunctionalization model were identified using the following criteria: a motif was considered if it was conserved over a region of at least 8 nt and lost in at least one paralog of the fish species for which multiple paralogs were present in the gene set. In order to minimize false positive motifs, extra constraints were set on the number of additional species in which the motif had to be conserved. Indeed, if conserved over larger phylogenetic distances, we can be more confident in the motif prediction. These constraints depended on the composition of the gene set:

If the gene set under study consisted of multiple non-fish homologs, either frog, chicken or human (which was the case for *bmp2*, *en2*, *glyRα*-related, *kcnip1*, *msx2*, *ntng1*, *ntng2*, *pax6*, *shh*, *six4*), a motif was only considered if it was conserved in at least two non-fish species. If the motif under study was derived from a gene set that contained only one non-fish homolog (*efna1* and *glyRα*), the motif had to be present in this one non-fish sequence.

For each motif we constructed a profile that indicates whether or not the motif occurs in the respective species from which the homologs of the gene family were derived. If the profile of the motif satisfies the requirements mentioned above, its profile was said to support

subfunctionalization. Phylogenetic profiles of motifs supporting subfunctionalization are represented in Table 3-3 and Table 3-4. To assess whether the detected motifs correspond to known transcription factor binding sites, we scanned the human instance of each conserved motif with the TRANSFAC 8.2 database of vertebrate transcription factor binding site profiles (Wingender et al. 2001). This scanning was performed using MotifLocator (Coessens et al. 2003; Aerts et al. 2003) with a 0th order vertebrate background model. Hits with a score >0.9 were regarded as potential binding sites. The binding sites are indicated by the Transfac factor name (Wingender et al. 2001). To further validate the link between the binding sites revealed with this screening and the gene under study, we did a text-based search with PubMed (2006b) using the name of the gene/protein under study (e.g. *pax6*) and the name of the transcription factor potentially binding the promoter region of this gene as search terms. When such a link existed, this is explicitly mentioned in the results section.

Chapter 6

Conclusions and perspectives

6.1 Conclusions

In this thesis we aimed at identifying novel transcription factor binding sites (regulatory motifs) in higher vertebrate genomes.

Motif detection algorithms provide one possibility to discover regulatory motifs. There applicability in higher vertebrates is, however, limited because of the long intergenic regions that are characteristic for vertebrate genes: this large amount of background sequence increases the signal-to-noise ratio and results in detection of many false positives, i.e., motifs with no biological function.

An alternative tool for discovering potential regulatory motifs is phylogenetic footprinting. This method uses cross-species comparisons to identify such regulatory motifs: it is based on the observation that functional elements evolve more slowly under selective pressure than non-functional sequences; by comparing intergenic sequences of orthologous genes, regulatory motifs can thus be identified as conserved 'islands' in a variable background sequence. But, due to the heterogeneity of vertebrate intergenics, the existing tools for phylogenetic footprinting are not optimally suited for motif discovery in higher vertebrates.

Since non of the existing tools for *de novo* motif discovery was appropriate for motif identification in strongly diverged vertebrates, in this thesis we developed novel strategies that perform better in this area of interest. The main **methodological contributions** of this thesis are (also see chapter 1: Figure 1-8 and Table 1-4):

- Development of a novel generic two-step procedure for phylogenetic footprinting that enables the identification of regulatory motifs in distantly related vertebrate genomes (chapter 2).

- Development of the first generic methodology to identify regulatory motifs in support of subfunctionalization on a genome wide scale (chapter 3).
- Development of a generic methodology to identify evolutionary conserved motifs that are shared by multiple co-expressed genes and thus possibly responsible for their similar expression pattern (chapter 5).

This thesis also describes the application of the developed methodologies on **biological *cases*** (also see chapter 1: Figure 1-8).

- Study of subfunctionalization:
  - Identification of regulatory motifs supporting experimentally observed expression divergence (chapter 3).
- Unravelling the mechanism of action of vitamin D_3:
  - Detection of known regulatory motifs in vitamin D_3-regulated genes using motif screening, leading to the identification of E2F as an important factor in the growth arrest induced by the active vitamin D_3 metabolite, 1,25-dihydroxyvitamin D_3 (vitD3) (chapter 4).
  - Identification of 31 regulatory motifs possibly involved in the molecular mechanism behind the action of vitD3 (chapter 5).

Chapter 2

- ⇨ We developed a two-step approach to identify evolutionary conserved regulatory motifs that combines the advantages of both motif detection and multiple alignment algorithms.
- ⇨ For the second step in this procedure, motif detection, we developed a new probabilistic algorithm, BlockSampler. This implementation identifies long conserved blocks in orthologous sequences.
- ⇨ The developed procedure proved to be well suited for identifying conserved regions in intergenic sequences from distantly related orthologs that show a low overall homology and that are heterogeneous in size.
- ⇨ The strength of our approach lies in the combination of data reduction and improved motif detection:

- The first data reduction step is essential when it concerns long intergenic sequences.
- BlockSampler, the algorithm used in the second motif detection step, is optimally suited to identify large conserved regions among orthologous sequences.

⇨ Application of our method to benchmark sets recovered most of the motifs previously described in these datasets.

⇨ Due to the stringent selection criteria applied some previously described motifs were missed.

⇨ Our method offers a fully automated analysis flow, which is highly specific in detecting motifs conserved over different vertebrate lineages in complete intergenic sequences.

Chapter 3

⇨ We applied the newly developed motif detection algorithm, BlockSampler, to several gene families, including families for which experimental evidence supported subfunctionalization (proof-of-concept).

⇨ For the majority of the proof-of-concept genes we identified motifs, which were differentially lost after a fish-specific duplication event and in accordance with the experimentally observed expression divergence.

⇨ Also for some additional gene families, we revealed motifs in support of subfunctionalization.

⇨ Because we applied very stringent selection criteria, in order to identify reliable regulatory motifs, we most likely underestimated the number of motifs indicative for subfunctionalization; relaxing the selection criteria will most likely result in the recovery of many more motifs supporting subfunctionalization.

Chapter 4

Besides the study of subfunctionalization, we focussed on a second biological topic, namely gaining more insight in the molecular mechanism behind the different effects -both classical and non-classical- of vitD3. Two different types of methodologies were applied to analyze sets of genes with a similar expression pattern after cells have been treated with vitD3: motif screening (chapter 4) and *de novo* motif detection (chapter 5).

- ⇨ First, we evaluated whether vitD3-induced expression patterns were the result of direct interaction of the vitamin D receptor, VDR, with its response element VDRE. Therefore, we screened the promoter regions of all genes under study with a motif model for VDRE. Our results indicate that VDR is not the main responsible for any of the observed expression patterns.
- ⇨ Next, we selected a cluster of down-regulated genes for further investigation.
    - The observation that many genes in this cluster are involved in cell cycle regulation and DNA replication, characterized them as potential players in the antiproliferative effects exerted by vitD3.
    - The fact that E2F is known to be a crucial effector in cell cycle progression, together with the observation that a large part of the genes in the selected cluster are established E2F targets, identified E2F as a potential effector in the vitD3-induced growth-inhibition.
    - We screened the promoter regions of the down-regulated genes with a motif model for the E2F binding site.
    - We discovered many additional E2F targets.
    - Two of novel E2F targets were selected for further wetlab experiments.
    - The complex between E2F and p107/p130 is probably involved in vitD3-induced growth-inhibition.

Chapter 5

- ⇨ We developed a novel procedure for *de novo* identification of regulatory motifs responsible for (at least part of) the expression pattern observed in co-expressed genes
- ⇨ This procedure combines two successful strategies, namely the two-step approach for phylogenetic footprinting (ortholog level) and motif detection in sets of co-expressed genes.
    - Ortholog level: for each individual co-expressed gene, we identified the motifs that were conserved during mammalian evolution.
    - Co-expression level: we searched for evolutionary conserved motifs shared by multiple co-expressed genes. Such motifs, present in the intergenic region of multiple

of newly developed tools in order to identify interesting alignment algorithms that can be used for data reduction.

- A second issue is the current implementation of BlockSampler. As most probabilistic motif detection tools, BlockSampler considers input sequences as statistically independent. Because its input consists of orthologous sequences this claim of independence is not fulfilled. Therefore, we are now working on an optimized algorithm that takes into account the phylogenetic relationship between the input sequences when calculating motif (block) scores: sequence conservation between highly diverged orthologs (e.g., human-fish) should have more importance than conservation between strongly related sequences (e.g., mammals). This can, for instance, be obtained by applying a weighting scheme that assigns a weight to each ortholog according to its position in the phylogenetic tree relating the input sequences.
- With the necessary technical improvements mentioned above, the improved two-step approach could be applied on a genome wide scale.
    - Since the developed methodology has proven successful for motif identification in distantly related vertebrate organisms, this could lead us to better insights of which gene functions have been retained during evolution. This would elucidate the underlying mechanisms that are common to all vertebrates.
    - Unlike in chapter 2, where we only considered motifs that were conserved in all species under investigation, it would also be interesting to assess which regulatory motifs are only retained in a vertebrate subgroup (for instance, mammals) and compare them with motifs conserved over longer periods of evolution (for instance mammals-fish, see chapter 2). This would teach us which regulatory mechanisms make us who we are (different levels possible: human vs. mammals, mammals vs. vertebrates) and which are common to a broader population of organisms.
- While in the work flow discussed in chapter 5, we first look for evolutionary conserved motifs for each separate co-expressed gene (orthologous data source) and then combine the motifs identified for all co-regulated genes in order to identify common motifs (co-expressed data source), we are now working on an implementation that will simultaneously consider orthologous and co-expression data.

- The comparison of multiple, moderately to strongly diverged species greatly improves motif detection by phylogenetic footprinting. Indeed, by aligning sequences that have undergone evolutionary changes, it becomes easier to distinct conserved functional elements, such as regulatory motifs, from non-conserved functionless DNA. Therefore, the strengths of our two-step approach for phylogenetic footprinting, which allows inclusion of multiple strongly divergent species, will become increasingly obvious as more and more vertebrate genomes will become publicly available.
- An obvious consequence of the developed phylogenetic footprinting approach is the possibility to discover completely new TFBSs. The experimental study of these binding sites can lead to elucidation of yet-unknown pathways. The fact that, unlike most methods for motif identification, the two-step procedure is able to identify regulatory motifs at long distances from the target gene will increase the amount of discovered motifs compared to methods that have a restricted search space.
- A very challenging future research direction is the genome wide study of subfunctionalization. Indeed, the results in chapter 3 indicate that subfunctionalization is likely to be more common in vertebrate evolution than generally assumed. Application of the developed procedure to all paralogous of an organism would yield an estimate of how many genes have changed function during evolution. Fish species make very interesting test cases: the great diversity of fish species (~ evolutionary success) is often assigned to the large number of genes in fish genome, which are the result of consecutive duplication events. We could evaluate this hypothesis by quantifying the number of changes in gene function in a fish genome that are due to sub- en neofunctionalization.
- The methodologies developed in this thesis can also be used to predict all the potential targets for every possible regulator protein. Using phylogenetic footprinting (two-step procedure) we are able to identify reliable regulatory motifs in sets of orthologous genes. These motifs can then be used to screen all the intergenic regions genome wide. In this way we can yield a compendium of all the genes that are potentially regulated by the same regulator. In the case of a known regulatory motif and/or protein, such a compendium contributes to clarifying the working mechanism of this specific regulator protein. On the other hand, if it concerns a regulatory motif that is recognized by an unknown regulator protein, a list of potential target genes of

this regulator make it more straightforward to deduce the protein's function: the molecular mechanism and function of the target genes will probably hint towards a possible functionality of the regulator.

- Because all the methods developed in this PhD are generic, the applications are endless:
  - As mentioned above, genome wide application would reveal large numbers of regulatory motifs, which would lead to a better understanding of gene regulatory networks and an improved comprehension of vertebrate functioning and evolution.
  - On the other hand, one might want to focus on a specific mechanism or pathway (e.g., vitD3-induced growth-inhibition). These would, for instance, be helpful in unravelling regulatory mechanisms underlying autoimmune diseases, cancer progression, ...

Appendix A

Additional results chapter 2

A.1 Introduction

This Appendix contains some additional results of the study described in chapter 2. In that chapter, we developed a novel methodology for identifying conserved motifs in sets of orthologous sequences (see Figure 2-1). This two-step procedure combines advantages of both motif detection and multiple alignment algorithms. We applied this two-step methodology to 10 datasets, i.e. four benchmark datasets and six additional datasets; details on their (initial) composition are given in §A.2.

The first step, data reduction, makes use of two algorithms, namely AVID (Bray et al. 2003) and TribeMCL (Enright et al. 2002) to select the conserved subsequences in intergenic orthologous sequences of comparable size of closely related vertebrate organisms. Such conserved subsequences are indeed expected to be rich in regulatory motifs (Woolfe et al. 2005). For more specific details on the data reduction step we refer to chapter 2, §2.5.2.1. The results of the data reduction step for the 10 datasets (§A.2) are then discussed in §A.3.1. In the second step, these preselected conserved sequences together with full-length intergenic sequence(s) of more distant ortholog(s) are subjected to motif detection. By applying BlockSampler we were able to identify long motifs that are conserved among different orthologous sequences: blocks. In §A.3.2 we give some additional information (compared to chapter 2) regarding the blocks recovered in each of the datasets.

As is explained in chapter 2, we evaluated the performance of our newly developed methodology by comparing its results on the benchmark datasets with the results of several other algorithms, both multiple alignment and motif detection algorithms (see §2.2.3). In §A.4 we take a closer look at the multiple alignments generated by the two-step methodology and MAVID (Bray and Pachter 2003).

Finally, in §A.5 we go further into detail on how the parameter for the different algorithms was chosen.

A.2 Input datasets

We applied the two-step procedure to four well-studied datasets for which functional phylogenetically conserved motifs had been extensively described: *cfos*, *hoxb2*, *pax6* and *scl*. These benchmark datasets always consist of orthologs of following species: human, chimp, mouse, rat and *Fugu* (Table A-1).

Table A-1. Detailed overview of the benchmark datasets. Gene: the name of the gene under study. Ensembl ID: the Ensembl Gene IDs of the different orthologs in human (ENSG), mouse (ENSMUSG), rat (ENSRNOG), chimp (ENSPTRG) and *Fugu rubripes* (SINFRUG) making up the benchmark datasets. Chrom: the chromosome on which the ortholog is located. Start G: the start of the gene as defined in Ensembl (www.ensembl.org; Birney et al. 2006). Stop G: end of the gene as defined in Ensembl. Strand: the strand on which the gene is located. 5'UTR: the stop position of the 5' UTR. Start I: the start of the selected intergenic region. Stop I: end position of the selected intergenic region. * The *Fugu scl* ortholog is the only exception for which the GenBank information is given: accession number.

Gene	Ensembl ID	Chrom	Start G	Stop G	Strand	5' UTR	Start I	Stop I
cfos	ENSG00000170345	14	73735572	73738948	1	73735726	73695572	73735726
	ENSMUSG00000021250	12	81466261	81469629	1	81466399	81433242	81466399
	ENSRNOG00000008015	6	109562610	109566002	1	109562742	109522610	109562742
	ENSPTRG00000006553	15	74847955	74851330	1	74848109	74807955	74848109
	SINFRUG00000132418	scaffold_164	262211	264241	-1	-	264241	267847
	SINFRUG00000132419	scaffold_164	267847	269581	1	-	264241	267847
	SINFRUG00000143787	scaffold_75	118455	120136	-1	-	120136	121380
hoxb2	ENSG00000173917	17	47094659	47097030	-1	47096912	47096912	47101885
	ENSMUSG00000047830	11	95934139	95937272	1	95936757	95930013	95936757
	ENSRNOG00000008365	10	85101118	85101603	1	-	85093478	85101118
	ENSPTRG00000009352	19_random	42925219	42925728	-1	-	42925728	42930606
	SINFRUG00000136637	scaffold_706	93771	96291	1	-	54552	93771
pax6	ENSG00000007372	11	31775791	31797074	-1	31796972	31796972	31837074
	ENSMUSG00000027168	2	105932344	105960811	1	-	105892344	105932344
	ENSRNOG00000004410	3	91023828	91045779	1	-	90983828	91023828
	ENSPTRG00000003474	9	32113516	32135710	-1	-	32135710	32175710
	SINFRUG00000121553	scaffold_227	2731	10658	-1	-	10658	31862
scl	ENSG00000162367	1	47051881	47065360	-1	47064785	47064785	47085766
	ENSMUSG00000028717	4	113968083	113980311	1	113968222	113951751	113968222
	ENSRNOG00000025051	5	135441839	135447104	1	-	135421496	135441839
	ENSPTRG00000000710	1	45596707	45610253	-1	45609676	45609676	45649675
*	AJ131019 (*Fugu rubripes*)							

Six additional datasets were included in the analysis: *egr3*, *gsh1*, *hiv-ep1*, *hoxb5*, *meis2* and *pcdh8*. These are sets of orthologous sequences for which, to our knowledge, no conserved motifs have previously been

reported. Furthermore, in contrast to the benchmark datasets, these additional datasets contain more than one distantly related organism (compared to mammals). They constitute of different combinations of human, chimp, mouse, rat, dog, chicken, *Fugu*, *Tetraodon* and zebrafish. The composition of these datasets is given in Table A-2.

Table A-2. Detailed overview of additional datasets. Legend see Table A-1. Chicken (ENSGALG), chimp (ENSPTRG), dog (ENSCAFG), *Fugu rubripes* (SINFRUG), human (ENSG), mouse (ENSMUSG), rat (ENSRNOG), and *Tetraodon nigroviridis* (GSTENG and HOXBb5).

Gene	Ensembl ID	Chrom	Start G	Stop G	Strand	5' UTR	Start I	Stop I
egr3	ENSG00000179388	8	22601119	22606760	-1	226/64/3	22606403	22626712
	ENSMUSG00000033730	14	61853137	61855849	1	-	61835571	61853137
	ENSPTRG00000020067	7	23537785	23543523	-1	23543134	23543134	23564160
	ENSCAFG00000009272	25	37622971	37625454	1	-	37605055	37622971
	GSTENG00020441001	12	712484	716154	-1	-	716154	756154
	ENSDARG00000011592	Zv4_scaffold105	60750	65223	1	6/968	20750	60968
gsh1	ENSG00000169840	13	27264780	27266089	1	27264827	27224780	27264827
	ENSMUSG00000053129	5	144513345	144514829	1	-	144473345	144513345
	ENSPTRG00000005735	14	26450716	26452026	1	2645/763	26410716	26450763
	ENSRNOG00000000952	12	8581078	8582373	1	-	8541078	8581078
	GSTENG00019041001	7	4250981	4251857	1	-	4243646	4250981
	SINFRUG00000149945	scaffold_6	404571	405414	1	-	405414	427424
hiv-ep1	ENSG00000095951	6	12120557	12273217	1	1212/785	12080557	12120785
	ENSMUSG00000021366	13	41502581	41631750	1	415/2581	41462581	41502581
	ENSPTRG00000017729	5	12400683	12553406	1	124//935	12360683	12400935
	ENSRNOG00000014460	17	28451133	28485482	-1	-	28485482	28525482
	ENSCAFG00000009791	35	14519321	14660786	1	-	14479321	14519321
	ENSGALG00000012739	2	62461406	62485723	-1	-	62485723	62525723
	GSTENG00007779001	Un_random	105144710	105177930	1	-	105122372	105144710
hoxb5	ENSG00000120075	17	44023619	44026117	-1	44/26/44	44026044	44028113
	ENSMUSG00000038700	11	95974635	95977243	1	95974736	95972689	95974736
	ENSPTRG00000009355	19_random	42972598	42975102	-1	42975/29	42975029	42977081
	ENSRNOG00000008010	10	85050005	85051489	1	-	85047334	85050005
	ENSGALG00000002996	Un	49806414	49807911	1	-	49775659	49806414
	HOXBb5	2	1432892	1434137	-1	-	1434137	1435393
	ENSDARG00000013057	19	29870866	29873261	-1	29873175	29873175	29875769
meis2	ENSG00000134138	15	34970525	35180796	-1	3518/385	35180385	35220796
	ENSMUSG00000027210	2	115644189	115846623	-1	11584639/	115846390	115886623
	ENSPTRG00000006901	16	34775125	34985957	-1	34984115	34984115	35025957
	ENSRNOG00000004730	3	101813557	102012150	-1	-	102012150	102052150
	ENSGALG00000009799	5	28113641	28275556	1	-	28073641	28113641
	GSTENG00032018001	10	7623006	7633708	1	-	7614977	7623006
	SINFRUG00000130030	scaffold_1302	25599	28989	1	-	17968	25599
	ENSDARG00000005196	17	42200185	42234131	1	-	42174719	42200185
pcdh8	ENSG00000136099	13	52316115	52320776	-1	5232/573	52320573	52360776
	ENSMUSG00000036422	14	71540299	71544814	-1	71544636	71544636	71584814
	ENSPTRG00000005909	14_random	23760367	23764828	1	-	23720367	23760367
	ENSRNOG00000013101	15	61072774	61076521	-1	-	61076521	61116521
	ENSGALG00000016944	1	157861150	157864657	1	-	157821150	157861150
	GSTENG00015083001	15	2450310	2453534	1	-	2438023	2450310
	SINFRUG00000153408	scaffold_1694	20220	23397	1	-	1	20220
	ENSDARG00000006467	Zv4_NA7075	19476	26189	1	19591	1	19591

A.3 Results of applying the two-step procedure: extra information

A.3.1 Data reduction

The data reduction procedure preselects subsequences conserved in closely related (mammalian) sequences. The resulting of data reduction are clusters consisting of a set of preselected sequences originating from different related orthologs of comparable size that mutually show a minimal degree of conservation. Such a cluster can thus be regarded as a local multiple alignment.

In our benchmark study, we were primarily interested in finding DNA motifs conserved among all input sequences (orthologs). Therefore, only clusters containing conserved subsequences of all mammalian orthologs included in this study (human, chimp, rat, and mouse) were retained for further analysis. The adequate clusters identified in the four benchmark datasets (*cfos*, *hoxb2*, *pax6* and *scl*) are specified in Table A-3.

It is possible that a cluster includes different subsequences originating from the same intergenic region, for instance when multiple smaller subsequences of a certain ortholog show homology with a single longer subsequence of another ortholog. This is the case for the cluster shown in Figure 2-4 (chapter 2): it contains a single homologous subsequence of each mammalian ortholog except for the rat for which the cluster contains two subsequences. Both rat subsequences show homology with a single region in the other three mammalian intergenic sequences, but are in the rat sequence separated by a non-conserved gap. Such gaps often result from sequences containing an NNN-stretch that does not match with the other intergenic sequences. Because these separated subsequences map to a different region of the intergenic region, they are not likely to contain the same motifs. Including them in the data set for motif detection would increase the noise in the motif detection input set. To minimize the noise in data sets used for motif detection, such clusters are split into subclusters (chapter 2, Figure 2-4). Each subcluster contains exactly one subsequence per ortholog and is represented by a profile consisting of the constituting subsequence ID's (one per ortholog), as given in Table A-3. The number of subclusters that have to be analyzed equals the number of possible combinations of region IDs, for which each ortholog is only included once. Table A-4 gives an overview of the generated subclusters for the benchmark datasets.

Each subcluster, together with the intergenic sequence of the corresponding *Fugu* ortholog, is then used as input for the subsequent motif detection step.

Table A-3. A detailed description of the selected clusters for the benchmark datasets, that is, the clusters that contain at least one sequence region of each orthologous intergenic sequence. Gene: the name of benchmark data set. Cluster: the cluster ID. Region ID: the region id within the cluster (used to generate a profile); Ensembl ID: the Ensembl gene ID from which this region is originating. Start: the start position of the region within the genome. Stop: the stop position of the region within the genome.

Gene	Cluster	Region ID	Ensembl ID	Start	Stop
cfos	1	6	ENSG00000170345	73732735	73735571
	1	7	ENSMUSG00000021250	81465628	81466260
	1	8	ENSRNOG00000008015	109561965	109562607
	1	9	ENSPTRG00000006553	74845113	74847954
	2	10	ENSG00000170345	73733540	73733841
	2	11	ENSMUSG00000021250	81463776	81465096
	2	12	ENSRNOG00000008015	109560141	109561506
	2	13	ENSPTRG00000006553	74845919	74846220
	3	14	ENSG00000170345	73731651	73731862
	3	15	ENSMUSG00000021250	81461044	81461513
	3	16	ENSRNOG00000008015	109557418	109557867
	3	17	ENSPTRG00000006553	74843941	74844152
hoxb2	0	1	ENSG00000173917	47100342	47097031
	0	2	ENSMUSG00000047830	95930013	95932973
	0	3	ENSMUSG00000047830	95933154	95934089
	0	4	ENSRNOG00000008365	85093478	85096407
	0	5	ENSRNOG00000008365	85096584	85097630
	0	6	ENSPTRG00000009352	42929041	42925728
	1	7	ENSG00000173917	47101885	47100420
	1	8	ENSMUSG00000047830	95930947	95931576
	1	9	ENSRNOG00000008365	85094470	85095030
	1	10	ENSPTRG00000009352	42930606	42929122
	2	11	ENSG00000173917	47097284	47097036
	2	12	ENSMUSG00000047830	95934862	95935404
	2	13	ENSRNOG00000008365	85098342	85098815
	2	14	ENSPTRG00000009352	42925985	42925728
pax6	0	1	ENSG00000007372	31809130	31796972
	0	2	ENSMUSG00000027168	105927848	105929364
	0	3	ENSRNOG00000004410	91012633	91014137
	0	4	ENSRNOG00000004410	91021300	91021525
	0	5	ENSRNOG00000004410	91023302	91023739
	0	6	ENSPTRG00000003474	32147753	32135710
	1	7	ENSG00000007372	31816169	31809195
	1	8	ENSMUSG00000027168	105924799	105925502
	1	9	ENSMUSG00000027168	105925609	105926706
	1	10	ENSRNOG00000004410	91009450	91010153
	1	11	ENSRNOG00000004410	91010387	91011487
	1	12	ENSPTRG00000003474	32154857	32147856
	2	13	ENSG00000007372	31812756	31812169
	2	14	ENSMUSG00000027168	105923278	105924708
	2	15	ENSRNOG00000004410	91007988	91009422
	2	16	ENSPTRG00000003474	32151423	32150866
scl	0	1	ENSG00000162367	47068148	47065361
	0	2	ENSMUSG00000028717	113964695	113967235
	0	3	ENSRNOG00000025051	135435196	135437737
	0	4	ENSPTRG00000000710	45613055	45610254

Table A-4. Overview of the relevant subclusters for each benchmark dataset, indicated by its profile, consisting of region IDs. Gene: the name of benchmark dataset; Cluster: the cluster ID; Profile: the subcluster profile, i.e., combination of region IDs (Table A-3), exactly one in each mammalian ortholog.

Gene	Cluster	Profile
cfos	1	6-7-8-9
	2	10-11-12-13
	3	14-15-16-17
hoxb2	1	1-2-4-6
		1-2-5-6
		1-3-4-6
		1-3-5-6
	2	7-8-9-10
	3	11-12-13-14
pax6	1	1-2-3-6
		1-2-4-6
		1-2-5-6
	2	7-8-10-12
		7-8-11-12
		7-9-10-12
		7-9-11-12
	3	13-14-15-16
scl	1	1-2-3-4

A.3.2 Motif detection results

A.3.2.1 Benchmark datasets

Using the two-step procedure we detected respectively 8 significant blocks for *hoxb2*, 13 for *pax6*, 1 in *scl* and none in the *cfos* data set (see chapter 2, Table 2-1). To validate these blocks we checked whether they contained transcription factor binding sites: we looked for previously described motifs (Kammandel et al. 1999; Gottgens et al. 2002; Scemama et al. 2002) and we also performed a screening with the TRANSFAC database of vertebrate transcription profiles (Wingender et al. 2001). The results are briefly summarized in tables 2-2, 2-3 and 2-4. A more detailed description of the regulatory motifs recovered in the detected blocks follows here for each benchmark dataset.

cfos

In none of the *cfos* datasets (Table A-1) we could detect significantly conserved blocks and thus the two (in pufferfish) conserved motifs previously described by Blanchette and Tompa (2002) could not be recovered. Our results indicate that the overall similarity between the mammalian and *Fugu cfos* orthologs is low and that the motifs, if biologically functional are not located in a conserved region. This hypothesis

is supported by the results when applying MAVID and TBA to each of the three cfos datasets: no conserved regions could be identified (see chapter 2, Table 2-1).

hoxb2

We identified eight significant blocks in the *hoxb2* data set (see chapter 2, Table 2-1). The location of these blocks on the complete intergenic region of the *hoxb2* ortholog of *Fugu* is shown in chapter 2 (Figure 2-2).

Block hoxb2 1.1 contains both the Hox/Pbx and Meis motifs previously reported by Scemama et al. (2002). TRANSFAC screening pointed out several additional potential binding sites in this block, e.g., a second Meis motif and several binding sites for homeodomain proteins. Scemama et al. (2002) also described some additional regulatory motifs located within this same region conserved between the vertebrate species. These motifs (Krox-20, Box1), shown to be essential for *hoxb2* expression in rhombomeres in mouse (Sham et al. 1993; Vesque et al. 1996; Maconochie et al. 1997), were not recovered by our methodology using the current selection criteria. Box1, for instance, was detected by BlockSampler, but in a very non-significant block (4th percentile). As was pointed out by Scemama et al. (2002), the location of the motifs box1 and Krox-20 in the different intergenic sequences is not conserved (i.e. upstream from the Meis/Hox/Pbx-location for mouse and downstream in *Fugu* and human). As a result, when aligning these motifs, the sequence surrounding these motifs might be less conserved. This explains why they remain undetected by our methodology.

Blocks hoxb2 2.1 to 2.4 contain motifs previously described by Scemama et al. (2002) (see chapter 2, Table 2-3), located in a region conserved in striped bass, zebrafish, mouse and human. At the time Scemama et al. (2002) performed the analysis the *Fugu* sequence was still incomplete. Therefore, they missed the corresponding region in the *Fugu hoxb2* intergenic region, which we identified in this study. As was to be expected, many of the motifs present in the striped bass *hoxb2a* intergenic region are also present in the closely related *Fugu* orthologous sequence (*Fugu* and striped bass diverged between 100 and 200 Mya). Screening with TRANSFAC pointed out the presence of additional potential motifs, some of which re-occur in more copies in the different blocks: for instance Cap, CdxA, NF-Y, SRY. Besides at the positions reported by Scemama et al. (2002), some of the previously described motifs, such as the octamer binding site and the CCAAT boxes, were also detected elsewhere in the (cluster 2) sequence (see chapter 2, Table 2-2). These repeated occurrences augment the confidence in the biological functionality of these detected motifs. Some motifs reported by Scemama et al. (2002) to be located within the same conserved region, however, could not be recovered by our methodology,

namely URTF, CBF1, Krox-20, Oct1, HNF1 and HOXD8,9,10. The latter three motifs, however, were detected by BlockSampler in cluster 2, but belonged to non-significant blocks (respectively the 4th, 18th and 39th percentile). Concerning URTF, CBF1 and Krox-20, the sensitivity of our methodology was too low to recover these motifs: URTF and CBF1 are lost in the preselection step, probably as a result of the chosen selection criteria. Krox-20 on the other hand is not present in the rat intergenic region and is thus lost in our analysis (see chapter 2, §2.5.2.2: selection procedure). Remark that block hoxb2 2.5 is located much further upstream of the transcription start site, circa 12 kb (see chapter 2, Figure 2-2), as compared to the location of the other detected blocks in the same cluster 2. According to Scemama et al. (2002) (based on *Fugu* genome consortium) the complete *Fugu hoxb2a* intergenic region comprises circa 5.4 kb only. This indicates that block hoxb2 2.5 is located outside this *hoxb2a* intergenic region. Closer inspection (using Ensembl) indicated that a pseudogene is present in the intergenic region of the *Fugu hoxb2* ortholog (see chapter 2, §2.5.1). The conserved block hoxb2 2.5 thus most likely corresponds to the regulatory region of this pseudogene.

Blocks 3.1 and 3.2 are significant blocks corresponding to conserved regions not previously described by Scemama et al. (2002). They are located near the transcription start site of the *Fugu hoxb2a* (see chapter 2, Figure 2-2) and contain some putative binding sites for upstream stimulating factors (Table 2-2: USF). TFIID, previously described by Scemama et al. (2002) was detected by BlockSampler, but in a non-significant block (77th percentile) located within cluster 3.

pax6

In the *pax6* data set, we detected 13 conserved blocks (see chapter 2, Table 2-1 and 2-3).

Six conserved blocks (chapter 2, Table 2-3: pax1.1-pax1.6) were located in mammalian cluster 1; their localization within the *Fugu* complete intergenic sequence is shown in chapter 2 (Figure 2-2). Block pax 1.1 (and Block pax 1.3 and 1.6) contains (a part of) the region described by Kammandel et al. (1999) as minimally required for expression in the lens and cornea (chapter 2, Table 2-3). Additionally, block pax 1.1 contains a motif with consensus "CTTAATG", also described by Kammandel et al. (1999). TRANSFAC screening identified latter motif as a homeobox-binding site (see chapter 2, Table 2-3: CdxA). Interestingly, many more CdxA binding sites were found in the *pax6* vertebrate sequences and they mostly occurred multiple times in the conserved blocks (chapter 2, Table 2-3). Block pax6 1.6 harbours many potential homeobox-binding sites, e.g., a homeodomain-binding site also reported by Kammandel et al. (1999) (see

chapter 2, Table 2-3). The motifs described above have been shown to be located in a region responsible for expression in eye tissues of head surface ectoderm origin (e.g. lens and cornea). In the pax1.5 block, located in same region (chapter 2, Figure 2-3), we identified a few not previously described binding sites such as SRY and EN-1 (see chapter 2, Table 2-3). Two blocks, pax6 1.2 and pax6 1.4, correspond to elements controlling expression in the developing pancreas described by Kammandel et al. (1999) (these are spatially separated from pax6 1.1, 1.3, 1.5 and 1.6 as is shown in chapter 2, Figure 2-2). These blocks are characterized by the presence of a PBX-1 binding site (pax6 1.2) and two motifs for homeodomain-binding sites (in respectively pax6 1.2 and 1.4) as was also previously described by Kammandel et al. (1999). TRANSFAC screening identified the homeodomain-binding site present in block pax6 1.2 as a HoxA3 motif (see chapter 2, Table 2-3).

In Cluster 2, we detected four not previously described conserved blocks: pax6 2.1 to pax6 2.4 (see chapter 2, Table 2-3). These blocks were rich in homeodomain binding sites such as CdxA, Nkx2-5, En-1. When looking at the localization of the identified cluster 2-blocks on the *pax6 Fugu* intergenic region (see chapter 2, Figure 2-2), it is remarkable that block pax6 2.4 is situated several kb downstream from the other blocks of that cluster, i.e., closer to the transcription start site. This can be due to the presence of a duplicated region within the *Fugu* intergenic sequence.

Finally, three significantly conserved blocks (see chapter 2, Table 2-3: pax3.1-pax3.3) were identified in *pax6* cluster 3 (see chapter 2, Figure 2-2). Also these blocks were not formerly described but contain many potential binding sites as identified by a screening with TRANSFAC. As was also observed in previous *pax6* conserved blocks, homeobox-binding sites, such as CdxA, En-1, Nkx2-5, Hoxa3 and Msx1 are abundantly present. Other binding sites that were identified multiple times in the different *pax6* blocks of cluster 3 are for instance the upstream stimulating factor and the sex-determining region Y product (SRY). Block pax6 3.3 also contains three GATA-binding sites.

scl

We detected one conserved block in the *scl* orthologs (see chapter 2, Table 2-4 and Figure 2-2). This block contained the conserved unnamed motifs and the putative SKN1 site formerly described by Göttgens et al. (2002). We could not identify the SKN1 by screening for existing motifs because TRANSFAC does not provide a vertebrate motif matrix for SKN. One of the unnamed motifs was identified by TRANSFAC screening as an En-1 or part of a HOXA3 binding site (see chapter 2, Table 2-4). As in the

previous data sets, such homeobox-binding sites were present abundantly. The two conserved GATA sites, formerly reported by Göttgens et al. (2002) were not detected by our methodology. These two motifs might either be too short or too isolated (i.e. no surrounding sequence conservation) to produce a significantly conserved block. For instance, the GATA site with consensus 'GCTTATCGGG' was recovered in a block with conservation level below our threshold (in the 42nd percentile).

A.3.2.2 Additional datasets

Table A-5. List of the significant blocks detected in the additional data sets. Block ID: the name of the identified block, which consists of the dataset name, the cluster number and a unique number. Consensus: the consensus sequence of the block. Organisms: the organisms the block is conserved in.

Block ID	Consensus	Organisms
egr3 1.1	TGnCnCGCnGCCCynCGACCCTCCCnnCA	human, chimp, mouse, dog, *Tetraodon*, zebrafish
egr3 1.2	TTGTCTGTCCATATATGGnCAnCTACGTCAC	human, chimp, mouse, dog, *Tetraodon*, zebrafish
gsh1 1.1	TnTTnCGGCGTGGGTGGGGnTGACAAGAATAGAnTACATTATGCAGTTCATTT AGTTAACAAGTGAAATAATGnGGAAGCGTGCAGnGnGAATGCCnAGAGAA	human, chimp, mouse, rat, *Fugu, Tetraodon*
gsh1 1.2	TTAGTTAACAAGTGAAATAATGnGGAAGCGTGCAGnGnGAATGCCnAGAGAAA nGnnnnAAAnnCnnTnnnG	human, chimp, mouse, rat, *Fugu, Tetraodon*
gsh1 1.3	AAAACCCTATTGAGAGnnnnnGGCCGCTnnnnGCGTAnn	human, chimp, mouse, rat, *Fugu, Tetraodon*
gsh1 1.4	AAAnTGAAAGAAAATGTTTTCCTATTACTTAATTCAATCAnAG	human, chimp, mouse, rat, *Fugu, Tetraodon*
hoxb5 1.1	GTCATnATTTGTAACCATAGAGCATGAATTACCTCTTGAnGTCATCAGnGAGAAT TTACGACTGGTCAACAAAnGCACGTGAT	human, chimp, mouse, rat, *Tetraodon*, zebrafish
hoxb5 1.2	CnCCCATATTTGGCCGCATACATAGCAAA	human, chimp, mouse, rat, *Tetraodon*, zebrafish
hoxb5 1.3	TAATTCATTAATACATCATAAATCGTGAAGCACAGGGTTATAACGACCAnGATC nACAAATCAAGCCCTCnAAAA	human, chimp, mouse, rat, *Tetraodon*, zebrafish
hoxb5 1.4	AAACGAAGTACAGTGCATnGCTATAATTCATTAATACATCATAAATCGTGAAG	human, chimp, mouse, rat, *Tetraodon*, zebrafish
hoxb5 1.5	AAACGAAGTACAGTGCATnGCTATAATTCATTAATACATCATA	human, chimp, mouse, rat, *Tetraodon*, zebrafish
hoxb5 1.6	ACGAAGTACAGTGCATnGCTATAATTCATTAATACATCATAAATCGTGA	human, chimp, mouse, rat, Tetraodon, zebrafish
meis2 1.1	TAACCCCAAAATGACCCCAATTTGACACCrCAAGATGAAATTTTACCGCCTGT TAAAACCA	human, chimp, mouse, rat, chicken, *Tetraodon*, zebrafish
meis2 1.2	ATGAAATTTTACCGCCTGTTAAAACCATTTCCAGCCTGGGCnn	human, chimp, mouse, rat, chicken, *Tetraodon*, zebrafish
pcdh8 1.1	CTAACGAGGGCTTCATAAGCCTTTGATACAGTCTGATCTTTGAAAC	human, chimp, mouse, rat, chicken, *Fugu, Tetraodon*

Six additional datasets were included in the analysis: *erg3*, *gsh1*, *hiv-ep1*, *hoxb5*, *meis2* and *pcdh8*. Each dataset consisted of different combinations of mammals chicken, *Fugu*, *Tetraodon* and zebrafish. The compositions of the initial datasets can be found in Table A-2. The detected (significant) blocks are enlisted in Table A-5. These blocks were screened with TRANSFAC (accordingly to the benchmark datasets); the results are summarized in Tables A-6 to A-10.

Erg3 is a developmental regulator (zinc finger protein), which is most likely involved in muscle spindle development. For the corresponding gene, two conserved blocks were found in mammals and fish (Table A-5 and Table A-6).

Table A-6. List of the significant blocks detected in the *egr3* data set. For each block, the consensus sequence is given, followed by the possible binding sites situated in this block; Motif hits derived by TRANSFAC are indicated by their matrix accession number, the consensus of this binding site and the instances of this motif in our search. These are further characterized by their positions relative to the consensus sequence of the entire block, by the strand (indicated by a "+" or a "-") on which the motif occurred and by the corresponding MotifLocator score (in parentheses).

Block	Consensus sequence and possible binding sites
egr3 1.1	**TGnCnCGCnGCCCynCGACCCTCCCnnCA**
egr3 1.2	**TTGTCTGTCCATATATGGnCAnCTACGTCAC**
	CdxA, M00101, AWTWMTR: 11-17 + (0,960); 10-16 – (0,960)
	CREB, M00039, TGACGTMA: 23-30 – (0,954)

Gsh1 has an important function in pituitary development. This data set contained four blocks that are conserved among human, chimp, mouse, rat, and pufferfish (*Fugu* and *Tetraodon*) (Table A-5 and Table A-7).

Hiv-ep1 is a zinc finger protein that specifically binds to enhancer elements present in several viral promoters and a number of cellular promoters such as those of the class I MHC, interleukin-2 receptor, and interferon-beta genes. No significant blocks were detected in this non-developmental gene.

Hoxb5 is part of a developmental regulatory system that provides cells with specific positional identities on the anterior-posterior axis. Six strongly conserved blocks were detected in *hoxb5* orthologs of mammals and pufferfish, while the motif seems to have been lost in chicken (Table A-5 and Table A-8).

Table A-7. List of the significant blocks detected in the *gsh1* data set. For legend see Table A-6.

Block	Consensus sequence and possible binding sites
gsh1 1.1	**TnTTnCGGCGTGGGTGGGGnTGACAAGAATAGAnTACATTATGCAGTTCATT**
	TAGTTAACAAGTGAAATAATGnGGAAGCGTGCAGnGnGAATGCCnAGAGAA
	Cap, M00253, NCANHNNN: 43-50 + (0,919); 48-55 + (0,908); 59-66 – (0,954)
	CdxA, M00100, MTTTATR: 37-43 + (0,917)
	CdxA, M00101, AWTWMTR: 37-43 + (0,929); 67-73 + (0,995), 53-59 – (0,917); 37-43 – (0.927)
	EGR2, M00246, NTGCGTRGGCGK: 6-17 + (0,912)
	EGR3, M00245, NTGCGTGGGCGK: 6-17 + (0,917)
	Nkx2-5, M00240, TYAAGTG: 59-65 + (0,934)
	SRY, M00148, AAACWAM: 52-58 – (0,922); 46-52 – (0,912)
gsh1 1.2	**TTAGTTAACAAGTGAAATAATGnGGAAGCGTGCAGnGnGAATGCCnAGAGAAAnGnnnnAAAnnCnnTnnnG**
	Cap, M00253, NCANHNNN: 8-15 – (0,954)
	CdxA, M00101, AWTWMTR: 16-22 + (0,995); 2-8 – (0,917)
	Nkx2-5, M00240, TYAAGTG: 8-14 + (0,934)
	SRY, M00148, AAACWAM: 1-7 – (0,922)
gsh1 1.3	**AAAACCCTATTGAGAGnnnnnGGCCGCTnnnnGCGTAnn**
gsh1 1.4	**AAAnTGAAAGAAAATGTTTTCCTATTACTTAATTCAATCAnAG**
	Cap, M00253, NCANHNNN: 10-17 – (0,919)
	CdxA, M00101, AWTWMTR: 24-30+ (0,923); 28-34 – (0,903); 23-29 – (0,921); 8-14 – (0,901)
	C/EBPalpha, M00116, NNATTRCNNAANNN: 22-35 + (0,905)
	C/EBPbeta, M00109, RNRTKDNGMAAKNN: 22-35 – (0,926)
	En-1, M00396, GTANTNN: 22-28 – (0,927)
	NF-AT, M00302, NANWGGAAAANN: 15-26 – (0,948)
	Nkx2-5, M00240, TYAAGTG: 26-32 – (0,911)
	Nkx2-5, M00241, CWTAATTG: 28-35 + (0,957)
	SRY, M00148, AAACWAM: 7-13 + (0,961)

Table A-8. List of the significant blocks detected in the *hoxb5* data set. For legend see Table A-6.

Block	Consensus sequence and possible binding sites
hoxb5 1.1	**GTCATnATTTGTAACCATAGAGCATGAATTACCTCTTGAnGT**
	CATCAGnGAGAATTTACGACTGGTCAACAAAnGCACGTGAT
	Cap, M00253, NCANHNNN: 58-65 – (0,949)
	CdxA, M00100, MTTTATR: 7-13 + (0,929); 54-60 + (0,941)
	CdxA, M00101, AWTWMTR: 54-60 + (0,958)
	En-1, M00396, GTANTNN: 26-32 – (0,950)
	HSF2 , M00147, NGAANNWTCK: 51-60 + (0,919)
	USF, M00217, NCACGTGN: 75-82 + (0,929); 75-82 – (0,982)
hoxb5 1.2	**CnCCCATATTTGGCCGCATACATAGCAAA**
	Cap, M00253, NCANHNNN: 16-23 + (0,901)
	CdxA, M00101, AWTWMTR: 5-11 – (0,953)
	En-1, M00396, GTANTNN: 6-12 + (0,915)
	HOXA3, M00395, CNTANNNKN: 5-13 + (0,948)
hoxb5 1.3	**TAATTCATTAATACATCATAAATCGTGAAGCACAGGGTTATAACGACCAnGATCnACAAATCAAGCCCTCnAAAA**
	Cap, M00253, NCANHNNN: 5-12 + (0,924); 16-23 + (0,901)
	CdxA, M00100, MTTTATR: 7-13 + (0,911); 7-13 – (1,00); 6-12 – (0,938)
	CdxA, M00101, AWTWMTR: 7-13 + (1,00); 39-45 + (0,925); 17-23 – (1,00); 6-12 – (1,00)
	Pbx-1, M00096, ANCAATCAW: 56-64 + (0,912)
	SRY, M00148, AAACWAM: 59-65 + (0,901)
Hoxb5 1.4	**AAACGAAGTACAGTGCATnGCTATAATTCATTAATACATCATAAATCGTGAAG**
	Cap, M00253, NCANHNNN: 28-35 + (0.924); 39-46 + (0,901)
	CdxA, M00100, MTTTATR: 30-36 + (0,911); 40-46 – (1,00); 29-35 – (0,938); 22-28 – (0,907)
	CdxA, M00101, AWTWMTR: 22-28 + (0,906); 30-36 + (1,00); 40-46 – (1,00); 29-35 – (1,00); 22-28 – (0,995)
	SRY, M00148, AAACWAM: 1-7 + (0,907)
hoxb5 1.5	**AAACGAAGTACAGTGCATnGCTATAATTCATTAATACATCATA**
	Cap, M00253, NCANHNNN: 28-35 + (0,924)
	CdxA, M00100, MTTTATR: 30-36 + (0,911); 29-35 – (0,938); 22-28 – (0,907);
	CdxA, M00101, AWTWMTR: 22-28 + (0,906);30-36 + (1,00); 29-35 – (1,00); 22-28 – (0,995)
	SRY, M00148, AAACWAM: 1-7 + (0,907)
hoxb5 1.6	**ACGAAGTACAGTGCATnGCTATAATTCATTAATACATCATAAATCGTGA**
	Cap, M00253, NCANHNNN: 26-33 + (0,924); 37-44 + (0,900)
	CdxA, M00100, MTTTATR: 28-34 + (0,911); 38-44 – (1,00); 27-33 – (0,938); 20-26 – (0,907)
	CdxA, M00101, AWTWMTR: 20-26 + (0,906); 28-34 + (1,00); 38-44 – (1,00); 27-33 – (1,00); 20-26 – (0,995)

meis2 is expressed in various tissues such as the lymphoid organs and some regions of the brain. For *meis2* two blocks were recovered that are retained in all organisms under study, except for *Fugu* (Table A-5 and Table A-9).

Table A-9. List of the significant blocks detected in the *meis2* data set. For legend see Table A-6.

Block	Consensus sequence and possible binding sites
meis2 1.1	**TAACCCCAAAATGACCCCAATTTGACACCrCAAGATGAAATTTTACCGCCTGTTAAAACCA**
	Cap, M00253, NCANHNNN: 31-38 – (0,930); 7-14 – (0,942)
	CdxA, M00101, AWTWMTR: 42-48 + (0,917); 51-57 – (0,930)
	RORalpha1, M00156, NWAWNNAGGTCAN: 11-23 – (0,911)
	TCF11, M00285, GTCATNNWNNNNN: 3-15 – (0,980)
meis2 1.2	**ATGAAATTTTACCGCCTGTTAAAACCATTTCCAGCCTGGGCnn**
	Cap, M00253, NCANHNNN: 25-32 + (0,915); 31-38 + (0,926);
	CdxA, M00101, AWTWMTR: 8-14 + (0,917); 27-33 + (0,942); 17-23 – (0.930)
	En-1, M00396, GTANTNN: 25-31 – (0,911)

The protocadherin 8 precursor Pcdh8 is believed to function as a calcium-dependent cell-adhesion protein. In this non-developmental related dataset one large block was detected, conserved in human, chimp, mouse, rat, chicken, *Tetraodon* and *Fugu*, but not in zebrafish (Table A-5 and Table A-10).

Table A-10. List of the significant blocks detected in the *pcdh8* data set. For legend see Table A-6.

Block	Consensus sequence and possible binding sites
pcdh 1.1	**CTAACGAGGGCTTCATAAGCCTTTGATACAGTCTGATCTTTGAAAC**
	Cap, M00253, NCANHNNN: 13-20 + (0,923); 28-35 + (0,936); 29-36 – (0,956)
	CdxA, M00101, AWTWMTR: 22-28 + (0,901)
	GATA-2, M00349, ASAGATAANA: 32-41 – (0,935)
	GATA-3, M00350, NGAGATAANA: : 32-41 – (0,918)
	GATA-3, M00351, ANAGATMWWA: 32-41 – (0,965)
	SRY, M00148, AAACWAM: 22-28 – (0,906)

A.4 Evaluation of the developed procedure: comparison MAVID and two-step procedure

In this paragraph, we will compare the (local) multiple alignments of the identified blocks generated by respectively the two-step procedure (Van Hellemont et al. 2005) and MAVID (Bray and Pachter 2003).

The detected blocks can be divided into blocks containing previously described motifs (§A.4.1) and novel, not previously reported conserved blocks (§A.4.2).

A.4.1 Blocks containing previously described motifs

For each benchmark datasets, we the alignments of the blocks that contain previously described motifs (see chapter 2, tables 2-2, 2-3 and 2-4) can be found on the supplementary website of the Genome Biology publication (Van Hellemont et al. 2005).

A.4.2 Newly identified conserved blocks

For each benchmark dataset, we show the alignments of the newly identified conserved blocks (see chapter 2, tables 2-2, 2-3 and 2-4) on the supplementary website of the Genome Biology publication (Van Hellemont et al. 2005).

A.5 Parameter settings

Our analysis flow consists of three major algorithms (AVID, §A.5.1; TribeMCL, §A.5.2; BlockSampler, §A.5.3), each attributed with its own set of parameters. Parameter fine-tuning of the major algorithms used in our analysis flow is based on multiple test runs with several benchmark data sets and different parameter settings. More information follows in this paragraph.

A.5.1 AVID parameters

Concerning the selection of pairwise conserved sequences (AVID-VISTA), the parameters L and C define the minimal length and degree of conservation for a region to be considered pairwise 'conserved' (Mayor et al. 2000; Bray et al. 2003; Frazer et al. 2004). Changing these parameter values

resulted in detecting the same (strongly conserved) regions. However, with lower parameter settings, instead of finding one large conserved region, the region was split into several smaller subpieces, matching the corresponding longer one. Different parameter settings were evaluated. The settings that resulted in the longest collinear regions were selected (see chapter 2, §2.5.2.1), since they reduce the complexity of the clustering in the subsequent step of the analysis flow.

A.5.2 TribeMCL parameters

For the clustering step, we evaluated the effect of the parameters I and P on the composition of the resulting clusters (Enright et al. 2002).

The inflation parameter I regulates the granularity and determines the size of the clusters (i.e. the number of sequences within the resulting clusters). This parameter did not have a major effect on the resulting clusters and was fixed at 4.

The other cluster parameter, P, is a similarity measure, determining the minimal similarity (i.e. the percent identity between two sequences) needed to group sequences together. To optimize this parameter, runs using different values of P were performed. The similarity measure P was adapted to the degree of phylogenetic relatedness between the organisms from which the pairwise compared sequences originated. Clustering was performed using either the same similarity threshold C as in the previous VISTA selection step (see chapter 2, §2.5.2.1), or by using less stringent criteria (subtracting 5%, respectively 10% from the conservation selection criterion C). For a given dataset, these tests thus resulted in three sets of clusters: one for threshold (P) score C%, one for C-5% and one for C-10% (as stated, I was kept constant at 4).

Dependent on the parameters used, a cluster set consisted of multiple small and tightly related clusters, or few large clusters. The large clusters contained weaker conserved subsequences, obtained by less stringent relations between the mammalian subsequences (e.g. lower percent identity). Because these clusters also contained subsequences that were not significantly homologous, the parameter set resulting in the smallest clusters was selected as input for motif detection. To determine the optimal setting of P, the quality (~size) of the clusters was evaluated for different values of P for each dataset separately. The parameters that resulted in the most tightly linked clusters for each dataset were chosen for motif detection; these are enumerated for each analyzed dataset in Table A-11.

Table A-11. Clustering parameters. Gene: the name of dataset. P: the similarity parameter and I: the inflation value parameter used to generate this set of clusters. Number of clusters: the number of clusters in the chosen set. Number of subsequences in the largest cluster.

Gene	P	I	Number of clusters	Number of subsequences in the largest cluster
cfos	0	4	12	5
hoxb2	-10	4	4	6
pax6	0	4	20	6
scl	-10	4	11	4
egr3	0	4	11	8
gsh1	-10	4	12	4
hiv-ep1	0	4	13	6
hoxb5	0	4	1	4
meis2	0	4	24	6
pchd8	0	4	14	4

A.5.3 BlockSampler

With the exception of the threshold on the consensus score, default parameter values were used to run the BlockSampler algorithm. The threshold itself was fixed at 1.2 to ensure that only initially well-conserved motifs were extended in length. Analysis that would require longer (shorter) and less (stronger) conserved blocks can be done by lowering (augmenting) this parameter value.

To select the most promising hits from the output of BlockSampler, we designed a score that is independent of block sequence length, but increases with the degree of conservation of the motifs. This normalized consensus score is appropriate because short motifs have a higher chance of resulting in a high consensus score. Normalization was done by recalculating the consensus scoring according to the formula $Cs_{ad} = (L/L+E).Cs$, where L is the length of the conserved block, E is an empirical factor and Cs is the consensus score. As a result, the normalized consensus score is not proportional to block sequence length. Different empirical factors were tested on different data sets, and 5 appeared to give the best balance between motif length and conservation. Again, depending on the interest of a particular study, the empirical factor can be enlarged to favor larger blocks.

References

1. Abe, E., Miyaura, C., Sakagami, H., Takeda, M., Konno, K., Yamazaki, T., Yoshiki, S., and Suda, T. 1981. Differentiation of mouse myeloid leukemia cells induced by 1 alpha,25-dihydroxyvitamin D3. Proc.Natl.Acad.Sci.U.S.A. 78:4990-4994
2. Abrahams, B.S., Mak, G.M., Berry, M.L., Palmquist, D.L., Saionz, J.R., Tay, A., Tan, Y.H., Brenner, S., Simpson, E.M., and Venkatesh, B. 2002. Novel vertebrate genes and putative regulatory elements identified at kidney disease and NR2E1/fierce loci. Genomics 80:45-53
3. Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y., and De Moor, B. 2003a. Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Res. 31:1753-1764
4. Aerts, S., Van Loo, P., Moreau, Y., and De Moor, B. 2004. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. Bioinformatics. 20:1974-1976
5. Aerts, S., Van Loo, P., Thijs, G., Mayer, H., de Martin, R., Moreau, Y., and De Moor, B. 2005. TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. Nucleic Acids Res. 33:W393-W396
6. Aerts, S., Van Loo, P., Thijs, G., Moreau, Y., and De Moor, B. 2003b. Computational detection of cis -regulatory modules. Bioinformatics. 19 Suppl 2:II5-II14
7. Alkema, W.B., Johansson, O., Lagergren, J., and Wasserman, W.W. 2004. MSCAN: identification of functional clusters of transcription factor binding sites. Nucleic Acids Res. 32:W195-W198
8. Altschmied, J., Delfgaauw, J., Wilde, B., Duschl, J., Bouneau, L., Volff, J.N., and Schartl, M. 2002. Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. Genetics. 161:259-267
9. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402
10. Anderson, K.P., Crable, S.C., and Lingrel, J.B. 1998. Multiple proteins binding to a GATA-E box-GATA motif regulate the erythroid Kruppel-like factor (EKLF) gene. J.Biol Chem. 273:14347-14354
11. Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297:1301-1310
12. Aparicio, S., Morrison, A., Gould, A., Gilthorpe, J., Chaudhuri, C., Rigby, P., Krumlauf, R., and Brenner, S. 1995. Detecting conserved regulatory

elements with the model genome of the Japanese puffer fish, Fugu rubripes. Proc.Natl.Acad.Sci.U.S.A 92:1684-1688

13. Arnone, M.I. and Davidson, E.H. 1997. The hardwiring of development: organization and function of genomic regulatory systems. Development. 124:1851-1864
14. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat.Genet. 25:25-29
15. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat.Genet. 25:25-29
16. Attwooll, C., Lazzerini, D.E., and Helin, K. 2004. The E2F family: specific functions and overlapping interests. EMBO J. 23:4709-4716
17. Bafna, V. and Huson, D.H. 2000. The conserved exon method for gene finding. Proc.Int.Conf.Intell.Syst.Mol.Biol. 8:3-12
18. Bagheri-Fam, S., Ferraz, C., Demaille, J., Scherer, G., and Pfeifer, D. 2001. Comparative genomics of the SOX9 region in human and Fugu rubripes: conservation of short regulatory sequence elements within large intergenic regions. Genomics 78:73-82
19. Bailey, T.L. and Elkan, C. 1995. The value of prior knowledge in discovering motifs with MEME. Proc.Int.Conf.Intell.Syst.Mol.Biol. 3:21-29
20. Bailey, T.L. and Noble, W.S. 2003. Searching for statistically significant regulatory modules. Bioinformatics.2003.Oct.;19.Suppl 2:II16.-II25. 19 Suppl 2:II16-II25
21. Ball, A.R., Jr., Schmiesing, J.A., Zhou, C., Gregson, H.C., Okada, Y., Doi, T., and Yokomori, K. 2002. Identification of a chromosome-targeting domain in the human condensin subunit CNAP1/hCAP-D2/Eg7. Mol.Cell Biol. 22:5769-5781
22. Balmer, J.E. and Blomhoff, R. 2006. Anecdotes, data and regulatory modules. Biol Lett. 2:431-434
23. Bamshad, M.J., Mummidi, S., Gonzalez, E., Ahuja, S.S., Dunn, D.M., Watkins, W.S., Wooding, S., Stone, A.C., Jorde, L.B., Weiss, R.B., et al. 2002. A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc.Natl.Acad.Sci.U.S A. 99:10539-10544
24. Bartek, J., Bartkova, J., and Lukas, J. 1996. The retinoblastoma protein pathway and the restriction point. Curr.Opin.Cell Biol. 8:805-814
25. Bartholin, L., Powers, S.E., Melhuish, T.A., Lasse, S., Weinstein, M., and Wotton, D. 2006. TGIF inhibits retinoid signaling. Mol.Cell Biol. 26:990-1001
26. Barton, L.M., Göttgens, B., Gering, M., Gilbert, J.G., Grafham, D., Rogers, J., Bentley, D., Patient, R., and Green, A.R. 2001. Regulation of the stem cell leukemia (SCL) gene: a tale of two fishes. Proc.Natl.Acad.Sci.U.S.A 98:6747-6752
27. Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: comparative analysis and application to exon prediction. Genome Res. 10:950-958

28. Bell, L.A. and Ryan, K.M. 2004. Life and death decisions by E2F-1. Cell Death.Differ. 11:137-142
29. Berezikov, E., Guryev, V., and Cuppen, E. 2005. CONREAL web server: identification and visualization of conserved transcription factor binding sites. Nucleic Acids Res. 33:W447-W450
30. Berezikov, E., Guryev, V., Plasterk, R.H., and Cuppen, E. 2004. CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. Genome Res. 14:170-178
31. Bertolino, E., Reimund, B., Wildt-Perinic, D., and Clerc, R.G. 1995. A novel homeobox protein which recognizes a TGT core and functionally interferes with a retinoid-responsive motif. J.Biol Chem. 270:31178-31188
32. Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., et al. 2006. Ensembl 2006. Nucleic Acids Res. 34:D556-D561
33. Blanchette, M. and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res. 12:739-748
34. Blanchette, M. and Tompa, M. 2003. FootPrinter: A program designed for phylogenetic footprinting. Nucleic Acids Res. 31:3840-3842
35. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14:708-715
36. Blanchette, M., Schwikowski, B., and Tompa, M. 2000. An exact algorithm to identify motifs in orthologous sequences from multiple species. Proc.Int.Conf.Intell.Syst.Mol.Biol. 8:37-45
37. BLAT Search Genome: http://genome.ucsc.edu/cgi-bin/hgBlat
38. Blayo, P., Rouze, P., and Sagot, M.-F. 2003. Orphan gene finding-an exon assembly approach. Theoret.Comput.Sci. 290:1407-1431
39. Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 7:R43.1-R43.12
40. Bockamp, E.O., McLaughlin, F., Murrell, A.M., Gottgens, B., Robb, L., Begley, C.G., and Green, A.R. 1995. Lineage-restricted regulation of the murine SCL/TAL-1 promoter. Blood 86:1502-1514
41. Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391-1394
42. Boffelli, D., Nobrega, M.A., and Rubin, E.M. 2004. Comparative genomics at the vertebrate extremes. Nat.Rev.Genet. 5:456-465
43. Bollig, F., Mehringer, R., Perner, B., Hartung, C., Schafer, M., Schartl, M., Volff, J.N., Winkler, C., and Englert, C. 2006. Identification and comparative expression analysis of a second wt1 gene in zebrafish. Dev Dyn. 235:554-561
44. Boonstra, A., Barrat, F.J., Crain, C., Heath, V.L., Savelkoul, H.F., and O'Garra, A. 2001. 1alpha,25-Dihydroxyvitamin d3 has a direct effect on

naive CD4(+) T cells to enhance the development of Th2 cells. J.Immunol 167:4974-4980

45. Bouillon, R., Verlinden, L., Eelen, G., De Clercq, P., Vandewalle, M., Mathieu, C., and Verstuyf, A. 2005. Mechanisms for the selective action of Vitamin D analogs. J.Steroid Biochem.Mol.Biol. 97:21-30
46. Boutros, P.C., Moffat, I.D., Franc, M.A., Tijet, N., Tuomisto, J., Pohjanvirta, R., and Okey, A.B. 2004. Dioxin-responsive AHRE-II gene battery: identification by phylogenetic footprinting. Biochem.Biophys.Res.Commun. 321:707-715
47. Bray, N. and Pachter, L. 2003. MAVID multiple alignment server. Nucleic Acids Res. 31:3525-3526
48. Bray, N. and Pachter, L. 2004. MAVID: constrained ancestral alignment of multiple sequences. Genome Res. 14:693-699
49. Bray, N., Dubchak, I., and Pachter, L. 2003. AVID: A global alignment program. Genome Res. 13:97-102
50. Brazma, A., Jonassen, I., Vilo, J., and Ukkonen, E. 1998. Predicting gene regulatory elements in silico on a genomic scale. Genome Res. 8:1202-1215
51. Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B., and Aparicio, S. 1993. Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature 366:265-268
52. Brudno, M. and Morgenstern, B. 2002. Fast and sensitive alignment of large genomic sequences. Proc.IEEE Comput.Soc.Bioinform.Conf. 1:138-147
53. Brudno, M., Chapman, M., Gottgens, B., Batzoglou, S., and Morgenstern, B. 2003a. Fast and sensitive multiple alignment of large genomic sequences. BMC.Bioinformatics. 4:66
54. Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. 2003b. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 13:721-731
55. Brudno, M., Poliakov, A., Salamov, A., Cooper, G.M., Sidow, A., Rubin, E.M., Solovyev, V., Batzoglou, S., and Dubchak, I. 2004. Automated whole-genome multiple alignment of rat, mouse, and human. Genome Res. 14:685-692
56. Brunet, A., Bonni, A., Zigmond, M.J., Lin, M.Z., Juo, P., Hu, L.S., Anderson, M.J., Arden, K.C., Blenis, J., and Greenberg, M.E. 1999. Akt promotes cell survival by phosphorylating and inhibiting a Forkhead transcription factor. Cell. 96:857-868
57. Bulyk, M.L. 2003. Computational prediction of transcription-factor binding site locations. Genome Biol.2003 5:201
58. Calhoun, V.C., Stathopoulos, A., and Levine, M. 2002. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. Proc.Natl.Acad.Sci.U.S A. 99:9243-9247
59. Cantorna, M.T., Hayes, C.E., and DeLuca, H.F. 1996. 1,25-Dihydroxyvitamin D3 reversibly blocks the progression of relapsing encephalomyelitis, a model of multiple sclerosis. Proc.Natl.Acad.Sci.U.S.A. 93:7861-7864

60. Cantorna, M.T., Hayes, C.E., and DeLuca, H.F. 1998. 1,25-Dihydroxycholecalciferol inhibits the progression of arthritis in murine models of human arthritis. J.Nutr. 128:68-72
61. Carlberg, C. 1995. Mechanisms of nuclear signalling by vitamin D3. Interplay with retinoid and thyroid hormone signalling. Eur.J.Biochem. 231:517-527
62. Carroll, S.B., Grenier, J.K., and Weatherbee, S.D. 2001. From DNA to diversity: molecular genetics and the evolution of animal design. Blackwell Science, Malden, Massachusetts, USA.
63. Casneuf, T., De Bodt, S., Raes, J., Maere, S., and Van de Peer Y. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. Genome Biol. 7:R13.1-R13.11
64. Chang, L., Khoo, B., Wong, L., and Tropepe, V. 2006. Genomic sequence and spatiotemporal expression comparison of zebrafish mbx1 and its paralog, mbx2. Dev Genes Evol. 216:647-654
65. Chapman, M.A., Donaldson, I.J., Gilbert, J., Grafham, D., Rogers, J., Green, A.R., and Gottgens, B. 2004. Analysis of multiple genomic sequence alignments: a web resource, online tools, and lessons learned from analysis of mammalian SCL loci. Genome Res. 14:313-318
66. Chellappan, S.P., Hiebert, S., Mudryj, M., Horowitz, J.M., and Nevins, J.R. 1991. The E2F transcription factor is a cellular target for the RB protein. Cell. 65:1053-1061
67. Chen, Q.K., Hertz, G.Z., and Stormo, G.D. 1995. MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. Comput.Appl.Biosci. 11:563-566
68. Chiaromonte, F., Weber, R.J., Roskin, K.M., Diekhans, M., Kent, W.J., and Haussler, D. 2003. The share of human genomic DNA under selection estimated from human-mouse genomic alignments. Cold Spring Harb.Symp.Quant.Biol. 68:245-254
69. Christakos, S., Dhawan, P., Shen, Q., Peng, X., Benn, B., and Zhong, Y. 2006. New insights into the mechanisms involved in the pleiotropic actions of 1,25dihydroxyvitamin D3. Ann.N.Y.Acad.Sci. 1068:194-203
70. Christensen, J., Cloos, P., Toftegaard, U., Klinkenberg, D., Bracken, A.P., Trinh, E., Heeran, M., Di Stefano, L., and Helin, K. 2005. Characterization of E2F8, a novel E2F-like cell-cycle regulated repressor of E2F-activated transcription. Nucleic Acids Res. 33:5458-5470
71. Christoffels, A., Koh, E.G., Chia, J.M., Brenner, S., Aparicio, S., and Venkatesh, B. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol.Biol Evol 21:1146-1151
72. Classon, M. and Harlow, E. 2002. The retinoblastoma tumour suppressor in development and cancer. Nat.Rev.Cancer. 2:910-917
73. Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301:71-76
74. Cliften, P.F., Hillier, L.W., Fulton, L., Graves, T., Miner, T., Gish, W.R., Waterston, R.H., and Johnston, M. 2001. Surveying Saccharomyces

genomes to identify functional elements by comparative DNA sequence analysis. Genome Res. 11:1175-1186

75. Coessens, B., Thijs, G., Aerts, S., Marchal, K., De Smet, F., Engelen, K., Glenisson, P., Moreau, Y., Mathys, J., and De Moor, B. 2003. INCLUSive: A web portal and service registry for microarray and regulatory sequence analysis. Nucleic Acids Res. 31:3468-3470
76. Davenne, M., Maconochie, M.K., Neun, R., Pattyn, A., Chambon, P., Krumlauf, R., and Rijli, F.M. 1999. Hoxa2 and Hoxb2 control dorsoventral patterns of neuronal development in the rostral hindbrain. Neuron 22:677-691
77. Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., et al. 2002. A genomic regulatory network for development. Science. 295:1669-1678
78. De Bodt, S., Theissen, G., and Van de Peer Y. 2006. Promoter Analysis of MADS-Box Genes in Eudicots Through Phylogenetic Footprinting. Mol.Biol.Evol.2006. 23:1293-1303
79. de Bruin, A., Maiti, B., Jakoi, L., Timmers, C., Buerki, R., and Leone, G. 2003. Identification and characterization of E2F7, a novel mammalian E2F family member capable of blocking cellular proliferation. J.Biol Chem. 278:42041-42049
80. Defrance, M. and Touzet, H. 2006. Predicting transcription factor binding sites using local over-representation and comparative genomics. BMC.Bioinformatics. 7:396-406
81. Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D.M., et al. 2002. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. S 298:2157-2167
82. Denny, P., Swift, S., Connor, F., and Ashworth, A. 1992. An SRY-related gene expressed during spermatogenesis in the mouse encodes a sequence-specific DNA-binding protein. EMBO J. 11:3705-3712
83. Dermitzakis, E.T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B.J., Flegel, V., Bucher, P., Jongeneel, C.V., et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature 2002.Dec.5;420.(6915.):578.-82. 420:578-582
84. Di Stefano, L., Jensen, M.R., and Helin, K. 2003. E2F7, a novel E2F featuring DP-independent repression of a subset of E2F-regulated genes. EMBO J. 22:6289-6298
85. DiLeone, R.J., Russell, L.B., and Kingsley, D.M. 1998. An extensive 3' regulatory region controls expression of Bmp5 in specific anatomical structures of the mouse embryo. Genetics. 148:401-408
86. Donaldson, I.J. and Gottgens, B. 2006. CoMoDis: composite motif discovery in mammalian genomes. Nucleic Acids Res. Electronic publication ahead of print
87. Down, T.A. and Hubbard, T.J. 2005. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. Nucleic Acids Res 33:1445-1453
88. Duret, L. and Bucher, P. 1997. Searching for regulatory elements in human noncoding sequences. Curr.Opin.Struct.Biol. 7:399-406

89. Dyson, N. 1998. The regulation of E2F by pRB-family proteins. Genes Dev. 12:2245-2262
90. E. Margulies ftp site: ftp://kronos.nhgri.nih.gov/pub/outgoing/elliott/tba/
91. ECR browser: http://ecrbrowser.dcode.org/
92. Eelen, G. Molecular modes of action of 1α,25-dihydroxyvitamin D3 and analogs. PhD thesis. Laboratorium voor Experimentele Geneeskunde en Endocrinologie, Faculteit Geneeskunde, Katholieke Universiteit Leuven, Herestraat 49, 3000 Leuven. 2005. Laboratorium voor Experimentele Geneeskunde en Endocrinologie, Faculteit Geneeskunde, Katholieke Universiteit Leuven, Herestraat 49, 3000 Leuven.
93. Eelen, G., Verlinden, L., Rochel, N., Claessens, F., De Clercq, P., Vandewalle, M., Tocchini-Valentini, G., Moras, D., Bouillon, R., and Verstuyf, A. 2005a. Superagonistic action of 14-epi-analogs of 1,25-dihydroxyvitamin D explained by vitamin D receptor-coactivator interaction. Mol.Pharmacol. 67:1566-1573
94. Eelen, G., Verlinden, L., Van Camp, M., Claessens, F., De Clercq, P., Vandewalle, M., Bouillon, R., and Verstuyf, A. 2005b. Altered Vitamin D receptor-coactivator interactions reflect superagonism of Vitamin D analogs. J.Steroid Biochem.Mol.Biol. 97:65-68
95. Eelen, G., Verlinden, L., Van Camp, M., Mathieu, C., Carmeliet, G., Bouillon, R., and Verstuyf, A. 2004a. Microarray analysis of 1alpha,25-dihydroxyvitamin D3-treated MC3T3-E1 cells. J.Steroid Biochem.Mol.Biol. 89-90:405-407
96. Eelen, G., Verlinden, L., Van Camp, M., Van Hummelen, P., Marchal, K., De Moor, B., Mathieu, C., Carmeliet, G., Bouillon, R., and Verstuyf, A. 2004b. The effects of 1alpha,25-dihydroxyvitamin D3 on the expression of DNA replication genes. J.Bone Miner.Res. 19:133-146
97. Ekker, M., Akimenko, M.A., Allende, M.L., Smith, R., Drouin, G., Langille, R.M., Weinberg, E.S., and Westerfield, M. 1997. Relationships among msx gene structure and function in zebrafish and other vertebrates. Mol.Biol.Evol. 14:1008-1022
98. Elemento, O. and Tavazoie, S. 2005. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol. 6:R18.1-R18.27
99. Elgar, G., Sandford, R., Aparicio, S., Macrae, A., Venkatesh, B., and Brenner, S. 1996. Small is beautiful: comparative genomics with the pufferfish (Fugu rubripes). Trends Genet. 12:145-150
100. Engelen, K., Coessens, B., Marchal, K., and De Moor, B. 2003. MARAN: normalizing micro-array data. Bioinformatics 19:893-894
101. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575-1584
102. Ensembl genome browser: http://www.ensembl.org
103. Eskin, E. and Pevzner, P.A. 2002. Finding composite regulatory patterns in DNA sequences. Bioinformatics. 18 Suppl 1:S354-S363
104. Eukaryotic Promoter Database (EPD): http://www.epd.isb-sib.ch.
105. Evans, M.J. and Scarpulla, R.C. 1990. NRF-1: a trans-activator of nuclear-encoded respiratory genes in animal cells. Genes Dev. 4:1023-1034

106. Felsenstein, J. PHYLIP -- Phylogeny Inference Package (Version 3.2). Cladistics 5, 164-166. 1989.
107. Fickett, J.W. and Wasserman, W.W. 2000. Discovery and modeling of transcriptional regulatory regions. Curr.Opin.Biotechnol. 11:19-24
108. Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151:1531-1545
109. Force, A., Shashikant, C., Stadler, P., and Amemiya, C.T. 2004. Comparative genomics, cis-regulatory elements, and gene duplication. Methods Cell Biol 77:545-561
110. Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., and Dubchak, I. 2004. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 32:W273-W279
111. Freeling, M. and Thomas, B.C. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 16:805-814
112. Frith, M.C., Hansen, U., and Weng, Z. 2001. Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics. 17:878-889
113. Frith, M.C., Li, M.C., and Weng, Z. 2003. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. Nucleic Acids Res. 31:3666-3668
114. Gaubatz, S., Wood, J.G., and Livingston, D.M. 1998. Unusual proliferation arrest and transcriptional control properties of a newly discovered E2F family member, E2F-6. Proc.Natl.Acad.Sci.U.S.A. 95:9190-9195
115. Gene Ontology Consortium: www.geneontology.org
116. Gilligan, P., Brenner, S., and Venkatesh, B. 2002. Fugu and human sequence comparison identifies novel human genes and conserved non-coding sequences. Gene 294:35-44
117. Givens, M.L., Kurotani, R., Rave-Harel, N., Miller, N.L., and Mellon, P.L. 2004. Phylogenetic footprinting reveals evolutionarily conserved regions of the gonadotropin-releasing hormone gene that enhance cell-specific expression. Mol.Endocrinol 18:2950-2966
118. GOLD (Genomes OnLine database): http://www.genomesonline.org
119. Gonzalez-Sancho, J.M., Larriba, M.J., Ordonez-Moran, P., Palmer, H.G., and Munoz, A. 2006. Effects of 1alpha,25-dihydroxyvitamin D3 in human colon cancer cells. Anticancer Res. 26:2669-2681
120. Gottgens, B., Barton, L.M., Chapman, M.A., Sinclair, A.M., Knudsen, B., Grafham, D., Gilbert, J.G., Rogers, J., Bentley, D.R., and Green, A.R. 2002. Transcriptional regulation of the stem cell leukemia gene (SCL)--comparative analysis of five vertebrate SCL loci. Genome Res. 12:749-759
121. Gottgens, B., Gilbert, J.G., Barton, L.M., Grafham, D., Rogers, J., Bentley, D.R., and Green, A.R. 2001. Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. Genome Res. 11:87-97
122. Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M., Rothberg, J.M., Paunovic, M., et al. 2006. Analysis of one million base pairs of Neanderthal DNA. Nature. 444:330-336

123. Gu, X. 2003. Evolution of duplicate genes versus genetic robustness against null mutations. Trends Genet. 19:354-356
124. Gu, X., Zhang, Z., and Huang, W. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc.Natl.Acad.Sci.U.S.A. 102:707-712
125. Gu, Z., Nicolae, D., Lu, H.H., and Li, W.H. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. 18:609-613
126. GuhaThakurta, D. 2006. Computational identification of transcriptional regulatory elements in DNA sequence. Nucleic Acids Res. 34:3585-3598
127. GuhaThakurta, D. and Stormo, G.D. 2001. Identifying target sites for cooperatively binding factors. Bioinformatics. 17:608-621
128. Gumucio, D.L., Heilstedt-Williamson, H., Gray, T.A., Tarle, S.A., Shelton, D.A., Tagle, D.A., Slightom, J.L., Goodman, M., and Collins, F.S. 1992. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. Mol.Cell Biol. 12:4919-4929
129. Gupta, M. and Liu, J.S. 2005. De novo cis-regulatory module elicitation for eukaryotic genomes. Proc.Natl.Acad.Sci.U.S A. 102:7079-7084
130. Hall, M., Bates, S., and Peters, G. 1995. Evidence for different modes of action of cyclin-dependent kinase inhibitors: p15 and p16 bind to kinases, p21 and p27 bind to cyclins. Oncogene. 11:1581-1588
131. Harper, J.W., Elledge, S.J., Keyomarsi, K., Dynlacht, B., Tsai, L.H., Zhang, P., Dobrowolski, S., Bai, C., Connell-Crowley, L., and Swindell, E. 1995. Inhibition of cyclin-dependent kinases by p21. Mol.Biol Cell. 6:387-400
132. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., et al. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32:D258-D261
133. Haubold, B. and Wiehe, T. 2004. Comparative genomics: methods and applications. Naturwissenschaften. 91:405-421
134. Haussler, M.R., Whitfield, G.K., Haussler, C.A., Hsieh, J.C., Thompson, P.D., Selznick, S.H., Dominguez, C.E., and Jurutka, P.W. 1998. The nuclear vitamin D receptor: biological and molecular regulatory properties revealed. J.Bone Miner.Res. 13:325-349
135. Hayes, C.E., Cantorna, M.T., and DeLuca, H.F. 1997. Vitamin D and multiple sclerosis. Proc.Soc.Exp.Biol Med. 216:21-27
136. He, X. and Zhang, J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169:1157-1164
137. Henikoff, S. and Henikoff, J.G. 1994. Position-based sequence weights. J.Mol.Biol. 243:574-578
138. Hertz, G.Z. and Stormo, G.D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15:563-577
139. Hertz, G.Z., Hartzell, G.W., III, and Stormo, G.D. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Comput.Appl.Biosci. 6:81-92

140. Heyer, B.S., Kochanowski, H., and Solter, D. 1999. Expression of Melk, a new protein kinase, during early mouse development. Dev Dyn. 215:344-351
141. Hirai, H., Roussel, M.F., Kato, J.Y., Ashmun, R.A., and Sherr, C.J. 1995. Novel INK4 proteins, p19 and p18, are specific inhibitors of the cyclin D-dependent kinases CDK4 and CDK6. Mol.Cell Biol. 15:2672-2681
142. Hogan, B.L. 1996. Bone morphogenetic proteins: multifunctional regulators of vertebrate development. Genes Dev. 10:1580-1594
143. Huang, H. and Tindall, D.J. 2006. FOXO factors: a matter of life and death. Future.Oncol. 2:83-89
144. Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J.Mol.Biol. 296:1205-1214
145. Imboden, M., Devignot, V., and Goblet, C. 2001. Phylogenetic relationships and chromosomal location of five distinct glycine receptor subunit genes in the teleost Danio rerio. Dev.Genes 211:415-422
146. INCLUSIVE motif finding tools: http://homes.esat.kuleuven.be/~thijs/download.html
147. Ingham, P.W. and McMahon, A.P. 2001. Hedgehog signaling in animal development: paradigms and principles. Genes Dev 15:3059-3087
148. International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 432:695-716
149. International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature.2001 409:860-921
150. Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946-957
151. Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res. 9:815-824
152. JASPAR database: http://mordor.cgb.ki.se/cgi-bin/jaspar2005/jaspar_db.pl
153. JGI: http://genome.jgi-psf.org
154. Jimenez-Delgado, S., Crespo, M., Permanyer, J., Garcia-Fernandez, J., and Manzanares, M. 2006. Evolutionary genomics of the recently duplicated amphioxus Hairy genes. Int.J.Biol.Sci. 2:66-72
155. Jochum, W., Passegue, E., and Wagner, E.F. 2001. AP-1 in mouse development and tumorigenesis. Oncogene 20:2401-2412
156. Johansson, O., Alkema, W., Wasserman, W.W., and Lagergren, J. 2003. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. Bioinformatics. 19 Suppl 1:i169-i176
157. Joyner, A.L. and Martin, G.R. 1987. En-1 and En-2, two mouse genes with sequence homolog to the Drosophila engrailed gene: expression during embryogenesis. Genes Dev 1:29-38
158. Kafri, R., Bar-Even, A., and Pilpel, Y. 2005. Transcription control reprogramming in genetic backup circuits. Nat.Genet. 37:295-299

159. Kahlen, J.P. and Carlberg, C. 1996. Functional characterization of a 1,25-dihydroxyvitamin D3 receptor binding site found in the rat atrial natriuretic factor promoter. Biochem Biophys.Res Commun. 218:882-886
160. Kammandel, B., Chowdhury, K., Stoykova, A., Aparicio, S., Brenner, S., and Gruss, P. 1999. Distinct cis-essential modules direct the time-space pattern of the Pax6 gene activity. Dev.Biol. 205:79-97
161. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC Genome Browser Database. Nucleic Acids Res 31:51-54
162. Kato, M., Hata, N., Banerjee, N., Futcher, B., and Zhang, M.Q. 2004. Identifying combinatorial regulation of transcription factors and binding motifs. Genome Biol. 5:R56.1-R56.13
163. Kellis, M., Patterson, N., Birren, B., Berger, B., and Lander, E.S. 2004. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. J.Comput.Biol. 11:319-355
164. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 423:241-254
165. Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. Genome Res 12:656-664
166. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. Genome Res 12:996-1006
167. Kirkness, E.F., Bafna, V., Halpern, A.L., Levy, S., Remington, K., Rusch, D.B., Delcher, A.L., Pop, M., Wang, W., Fraser, C.M., et al. 2003. The dog genome: survey sequencing and comparative analysis. Science. 301:1898-1903
168. Kliewer, S.A., Umesono, K., Noonan, D.J., Heyman, R.A., and Evans, R.M. 1992. Convergence of 9-cis retinoic acid and peroxisome proliferator signalling pathways through heterodimer formation of their receptors. Nature. 358:771-774
169. Kops, G.J., Dansen, T.B., Polderman, P.E., Saarloos, I., Wirtz, K.W., Coffer, P.J., Huang, T.T., Bos, J.L., Medema, R.H., and Burgering, B.M. 2002a. Forkhead transcription factor FOXO3a protects quiescent cells from oxidative stress. Nature. 419:316-321
170. Kops, G.J., Medema, R.H., Glassford, J., Essers, M.A., Dijkers, P.F., Coffer, P.J., Lam, E.W., and Burgering, B.M. 2002b. Control of cell cycle exit and entry by protein kinase B-regulated forkhead transcription factors. Mol.Cell Biol. 22:2025-2036
171. Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. Bioinformatics. 17 Suppl 1:S140-S148
172. Kratochwil, K., Dull, M., Farinas, I., Galceran, J., and Grosschedl, R. 1996. Lef1 expression is activated by BMP-4 and regulates inductive tissue interactions in tooth and hair development. Genes Dev. 10:1382-1394
173. Krivan, W. and Wasserman, W.W. 2001. A predictive model for regulatory sequences directing liver-specific transcription. Genome Res. 11:1559-1566

174. Kuras, L. and Struhl, K. 1999. Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme. Nature. 399:609-613
175. Laforest, L., Brown, C.W., Poleo, G., Geraudie, J., Tada, M., Ekker, M., and Akimenko, M.A. 1998. Involvement of the sonic hedgehog, patched 1 and bmp2 genes in patterning of the zebrafish dermal fin rays. Development. 125:4175-4184
176. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. Nature.2001 409:860-921
177. Latchman, D.S. 1998. Eukaryotic transcription factors. Academic Press, San Diego, California, USA.
178. Lavia, P. and Jansen-Durr, P. 1999. E2F target genes and cell-cycle checkpoint control. Bioessays. 21:221-230
179. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262:208-214
180. Lee, T.I. and Young, R.A. 2000. Transcription of eukaryotic protein-coding genes. Annu.Rev.Genet. 34:77-137
181. Lemon, B. and Tjian, R. 2000. Orchestrated response: a symphony of transcription factors for gene control. Genes Dev. 14:2551-2569
182. Lemos, B., Yunes, J.A., Vargas, F.R., Moreira, M.A., Cardoso, A.A., and Seuanez, H.N. 2004. Phylogenetic footprinting reveals extensive conservation of Sonic Hedgehog (SHH) regulatory elements. Genomics. 84:511-523
183. Lenhard, B. and Wasserman, W.W. 2002. TFBS: Computational framework for transcription factor binding site analysis. Bioinformatics. 18:1135-1136
184. Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., Van de, P.Y., Rouze, P., and Rombauts, S. 2002. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res. 30:325-327
185. Levy, S., Hannenhalli, S., and Workman, C. 2001. Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics 17:871-877
186. Liolios, K., Tavernarakis, N., Hugenholtz, P., and Kyrpides, N.C. 2006. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res. 34:D332-D334
187. Liu, X., Brutlag, D.L., and Liu, J.S. 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac.Symp.Biocomput.:127-138
188. Lo, R.S., Wotton, D., and Massague, J. 2001. Epidermal growth factor signaling via Ras controls the Smad transcriptional co-repressor TGIF. EMBO J. 20:128-136
189. Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science. 288:136-140

190. Loots, G.G., Ovcharenko, I., Pachter, L., Dubchak, I., and Rubin, E.M. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res. 12:832-839
191. Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155
192. Lynch, M. and Force, A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459-473
193. Maconochie, M.K., Nonchev, S., Studer, M., Chan, S.K., Popperl, H., Sham, M.H., Mann, R.S., and Krumlauf, R. 1997. Cross-regulation in the mouse HoxB complex: the expression of Hoxb2 in rhombomere 4 is regulated by Hoxb1. Genes Dev. 11:1885-1895
194. Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. Proc.Natl.Acad.Sci.U.S.A. 102:5454-5459
195. Major, M.B. and Jones, D.A. 2004. Identification of a gadd45beta 3' enhancer that mediates SMAD3- and SMAD4-dependent transcriptional induction by transforming growth factor beta. J.Biol.Chem. 279:5278-5287
196. Mar, L. and Hoodless, P.A. 2006. Embryonic fibroblasts from mice lacking Tgif were defective in cell cycling. Mol.Cell Biol. 26:4302-4310
197. Marchal, K., De Keersmaecker, S., Monsieurs, P., van Boxel, N., Lemmens, K., Thijs, G., Vanderleyden, J., and De Moor, B. 2004. In silico identification and experimental validation of PmrAB targets in Salmonella typhimurium by regulatory motif detection. Genome Biol. 5:R9.1-R9.20
198. Marchal, K., Thijs, G., De Keersmaecker, S., Monsieurs, P., De Moor, B., and Vanderleyden, J. 2003. Genome-specific higher-order background models to improve motif detection. Trends Microbiol. 11:61-66
199. Margulies, E.H., Blanchette, M., Haussler, D., and Green, E.D. 2003. Identification and characterization of multi-species conserved sequences. Genome Res. 13:2507-2518
200. Markey, M., Siddiqui, H., and Knudsen, E.S. 2004. Geminin is targeted for repression by the retinoblastoma tumor suppressor pathway through intragenic E2F sites. J.Biol Chem. 279:29255-29262
201. Martinez-Barbera, J.P., Toresson, H., Da Rocha, S., and Krauss, S. 1997. Cloning and expression of three members of the zebrafish Bmp family: Bmp2a, Bmp2b and Bmp4. Gene. 198:53-59
202. Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 31:374-378
203. MAVID multiple alignment server: http://baboon.math.berkeley.edu/mavid/
204. Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics. 16:1046-1047
205. Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics. 16:1046-1047

206. McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V., and Lawrence, C.E. 2001. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. Nucleic Acids Res. 29:774-782
207. McGuire, A.M., Hughes, J.D., and Church, G.M. 2000. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. Genome Res. 10:744-757
208. McIlhatton, M.A., Bremner, P., McMullin, M.F., Maxwell, A.P., Winter, P.C., and Lappin, T.R. 1998. Sequence characterisation and expression of homeobox HOX A7 in the multi-potential erythroleukaemic cell line TF-1. Biochim.Biophys.Acta. 1442:329-333
209. Medema, R.H., Kops, G.J., Bos, J.L., and Burgering, B.M. 2000. AFX-like Forkhead transcription factors mediate cell-cycle regulation by Ras and PKB through p27kip1. Nature 404:782-787
210. Meyer, A. and Van de Peer Y. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). BioEssays 27:937-945
211. Meyer, I.M. and Durbin, R. 2002. Comparative ab initio prediction of gene structures using pair HMMs. Bioinformatics. 18:1309-1318
212. Meyer, I.M. and Durbin, R. 2004. Gene structure conservation aids similarity based gene prediction. Nucleic Acids Res.2004 Feb.4;32.(2):776.-83.Print.2004. 32:776-783
213. MGI database: www.informatics.jax.org
214. Miller Lab: http://bio.cse.psu.edu/
215. Mittnacht, S. 1998. Control of pRB phosphorylation. Curr.Opin.Genet Dev. 8:21-27
216. Monsieurs, P., Thijs, G., Fadda, A.A., De Keersmaecker, S.C., Vanderleyden, J., De Moor, B., and Marchal, K. 2006. More robust detection of motifs in coexpressed genes by using phylogenetic information. BMC.Bioinformatics. 7:160
217. Montpetit, A. and Sinnett, D. 2001. Comparative analysis of the ETV6 gene in vertebrate genomes from pufferfish to human. Oncogene 20:3437-3442
218. Moore, R.C. and Purugganan, M.D. 2005. The evolutionary dynamics of plant duplicate genes. Curr.Opin.Plant Biol. 8:122-128
219. Morgenstern, B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15:211-218
220. Morgenstern, B., Frech, K., Dress, A., and Werner, T. 1998. DIALIGN: finding local similarities by multiple sequence alignment. Bioinformatics 14:290-294
221. Morgenstern, B., Rinner, O., Abdeddaim, S., Haase, D., Mayer, K.F., Dress, A.W., and Mewes, H.W. 2002. Exon discovery by genomic sequence alignment. Bioinformatics. 18:777-787
222. Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520-562
223. Muller, F., Blader, P., and Strahle, U. 2002. Search for enhancers: teleost models in comparative genomic and transgenic analysis of cis regulatory elements. Bioessays. 24:564-572

224. National Centre for Biotechnology Information (NCBI): http://www.ncbi.nlm.nih.gov/
225. Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J.Mol.Biol. 48:443-453
226. Nelson, C.E., Hersh, B.M., and Carroll, S.B. 2004. The regulatory content of intergenic DNA shapes genome architecture. Genome Biol. 5:R25
227. Nevins, J.R. 1992. E2F: a link between the Rb tumor suppressor protein and viral oncoproteins. Science. 258:424-429
228. Neznanov, N., Umezawa, A., and Oshima, R.G. 1997. A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice. J.Biol Chem. 272:27549-27557
229. NIH Intramural Sequencing Center (NISC): http://www.nisc.nih.gov/
230. Nishimura, S., Takahashi, S., Kuroha, T., Suwabe, N., Nagasawa, T., Trainor, C., and Yamamoto, M. 2000. A GATA box in the GATA-1 gene hematopoietic enhancer is a critical element in the network of GATA factors and sites that regulate this gene. Mol.Cell Biol. 20:713-723
231. Nobrega, M.A., Ovcharenko, I., Afzal, V., and Rubin, E.M. 2003. Scanning human gene deserts for long-range enhancers. Science 302:413
232. Nomenclature Committee for the International Union of Biochemistry (NC-IUB). 1986. Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Mol.Biol Evol. 3:99-108
233. Noonan, J.P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Paabo, S., Pritchard, J.K., et al. 2006. Sequencing and analysis of Neanderthal genomic DNA. Science. 314:1113-1118
234. Nornes, S., Clarkson, M., Mikkola, I., Pedersen, M., Bardsley, A., Martinez, J.P., Krauss, S., and Johansen, T. 1998. Zebrafish contains two pax6 genes involved in eye development. Mech.Dev. 77:185-196
235. Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J.Mol.Biol. 302:205-217
236. Novichkov, P.S., Gelfand, M.S., and Mironov, A.A. 2001. Gene recognition in eukaryotic DNA by comparison of genomic sequences. Bioinformatics. 17:1011-1018
237. Ogawa, H., Ishiguro, K., Gaubatz, S., Livingston, D.M., and Nakatani, Y. 2002. A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells. Science 296:1132-1136
238. Osada, H., Grutz, G.G., Axelson, H., Forster, A., and Rabbitts, T.H. 1997. LIM-only protein Lmo2 forms a protein complex with erythroid transcription factor GATA-1. Leukemia. 11 Suppl 3:307-312
239. Palmer, H.G., Gonzalez-Sancho, J.M., Espada, J., Berciano, M.T., Puig, I., Baulida, J., Quintanilla, M., Cano, A., de Herreros, A.G., Lafarga, M., et al. 2001. Vitamin D(3) promotes the differentiation of colon carcinoma cells by the induction of E-cadherin and the inhibition of beta-catenin signaling. J.Cell Biol. 154:369-387
240. Papp, B., Pal, C., and Hurst, L.D. 2003. Evolution of cis-regulatory elements in duplicated genes of yeast. Trends Genet. 19:417-422

241. Parra, G., Agarwal, P., Abril, J.F., Wiehe, T., Fickett, J.W., and Guigo, R. 2003. Comparative gene prediction in human and mouse. Genome Res. 13:108-117
242. Patti, M.E. 2004. Gene expression in humans with diabetes and prediabetes: what have we learned about diabetes pathophysiology? Curr.Opin.Clin.Nutr.Metab Care 7:383-390
243. Pavesi, G., Mauri, G., and Pesole, G. 2004a. In silico representation and discovery of transcription factor binding sites. Brief.Bioinform. 5:217-236
244. Pavesi, G., Mereghetti, P., Mauri, G., and Pesole, G. 2004b. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. 32:W199-W203
245. Pavesi, G., Mereghetti, P., Zambelli, F., Stefani, M., Mauri, G., and Pesole, G. 2006. MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. Nucleic Acids Res. 34:W566-W570
246. Pennacchio, L.A. 2003. Insights from human/mouse genome comparisons. Mamm.Genome 14:429-436
247. Pevzner, P.A. and Sze, S.H. 2000. Combinatorial approaches to finding subtle signals in DNA sequences. Proc.Int.Conf.Intell.Syst.Mol.Biol. 8:269-278
248. Philippakis, A.A., He, F.S., and Bulyk, M.L. 2005. Modulefinder: a tool for computational discovery of cis regulatory modules. Pac.Symp.Biocomput.:519-530
249. PlantCARE: http://bioinformatics.psb.ugent.be/webtools/plantcare/html/
250. Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E., and Eisen, M.B. 2004. Benchmarking tools for the alignment of functional noncoding DNA. BMC.Bioinformatics. 5:6-22
251. Polyak, K., Kato, J.Y., Solomon, M.J., Sherr, C.J., Massague, J., Roberts, J.M., and Koff, A. 1994a. p27Kip1, a cyclin-Cdk inhibitor, links transforming growth factor-beta and contact inhibition to cell cycle arrest. Genes Dev. 8:9-22
252. Polyak, K., Lee, M.H., Erdjument-Bromage, H., Koff, A., Roberts, J.M., Tempst, P., and Massague, J. 1994b. Cloning of p27Kip1, a cyclin-dependent kinase inhibitor and a potential mediator of extracellular antimitogenic signals. Cell. 78:59-66
253. Postlethwait, J., Amores, A., Cresko, W., Singer, A., and Yan, Y.L. 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends Genet 20:481-490
254. Postlethwait, J.H., Yan, Y.L., Gates, M.A., Horne, S., Amores, A., Brownlie, A., Donovan, A., Egan, E.S., Force, A., Gong, Z., et al. 1998. Vertebrate genome evolution and the zebrafish gene map. Nat.Genet. 18:345-349
255. Praz, V., Perier, R., Bonnard, C., and Bucher, P. 2002. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. Nucleic Acids Res. 30:322-324
256. Prestridge, D.S. 1991. SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements. Comput.Appl.Biosci. 7:203-206

257. Prestridge, D.S. 1996. SIGNAL SCAN 4.0: additional databases and sequence formats. Comput.Appl.Biosci. 12:157-160
258. Prince, V.E. and Pickett, F.B. 2002. Splitting pairs: the diverging fates of duplicated genes. Nat.Rev.Genet. 3:827-837
259. Promoter database of Saccharomyces cerevisiae (SCPD): http://rulai.cshl.edu/SCPD/
260. PubMed: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
261. Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. 23:4878-4884
262. Rabbits, T.H. 1991. Translocations, master genes, and differences between the origins of acute and chronic leukemias. Cell, 67: 641-644
263. Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 428:493-521
264. Ray, A. and Ray, B.K. 1998. Isolation and functional characterization of cDNA of serum amyloid A-activating factor that binds to the serum amyloid A promoter. Mol.Cell Biol. 18:7327-7335
265. Ray, A., Kumar, D., Shakya, A., Brown, C.R., Cook, J.L., and Ray, B.K. 2004a. Serum amyloid A-activating factor-1 (SAF-1) transgenic mice are prone to develop a severe form of inflammation-induced arthritis. J.Immunol. 173:4684-4691
266. Ray, A., Kuroki, K., Cook, J.L., Bal, B.S., Kenter, K., Aust, G., and Ray, B.K. 2003. Induction of matrix metalloproteinase 1 gene expression is regulated by inflammation-responsive transcription factor SAF-1 in osteoarthritis. Arthritis Rheum. 48:134-145
267. Ray, A., Shakya, A., Kumar, D., and Ray, B.K. 2004b. Overexpression of serum amyloid A-activating factor 1 inhibits cell proliferation by the induction of cyclin-dependent protein kinase inhibitor p21WAF-1/Cip-1/Sdi-1 expression. J.Immunol. 172:5006-5015
268. Rinner, O. and Morgenstern, B. 2002. AGenDA: gene prediction by comparative sequence analysis. In Silico.Biol. 2:195-205
269. Rodriguez-Trelles, F., Tarrio, R., and Ayala, F.J. 2003. Evolution of cis-regulatory regions versus codifying regions. Int.J.Dev Biol. 47:665-673
270. Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouze, P., and Van de, P.Y. 2003. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol. 132:1162-1176
271. Roose, J. and Clevers, H. 1999. TCF transcription factors: molecular switches in carcinogenesis. Biochim.Biophys.Acta. 1424:M23-M37
272. Roose, J., Huls, G., van Beest, M., Moerer, P., van der, H.K., Goldschmeding, R., Logtenberg, T., and Clevers, H. 1999. Synergy between tumor suppressor APC and the beta-catenin-Tcf4 target Tcf1. Science. 285:1923-1926
273. Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat.Biotechnol. 16:939-945

274. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 32:D91-D94
275. Sandelin, A., Bailey, P., Bruce, S., Engstrom, P.G., Klos, J.M., Wasserman, W.W., Ericson, J., and Lenhard, B. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC.Genomics. 5:99-107
276. Sandve, G.K. and Drablos, F. 2006. A survey of motif discovery methods in an integrated framework. Biol Direct. 1:11
277. Santini, S., Boore, J.L., and Meyer, A. 2003. Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. Genome Res. 13:1111-1122
278. Scemama, J.L., Hunter, M., McCallum, J., Prince, V., and Stellwag, E. 2002. Evolutionary divergence of vertebrate Hoxb2 expression patterns and transcriptional regulatory loci. J.Exp.Zool. 294:285-299
279. Schafer, K.A. 1998. The cell cycle: a review. Vet.Pathol. 35:461-478
280. Schmidt, M., Fernandez, d.M., van der, H.A., Klompmaker, R., Kops, G.J., Lam, E.W., Burgering, B.M., and Medema, R.H. 2002. Cell cycle inhibition by FoxO forkhead transcription factors involves downregulation of cyclin D. Mol.Cell Biol. 22:7842-7852
281. Schneider, T.D., Stormo, G.D., Gold, L., and Ehrenfeucht, A. 1986. Information content of binding sites on nucleotide sequences. J.Mol.Biol. 188:415-431
282. Schrader, M., Nayeri, S., Kahlen, J.P., Muller, K.M., and Carlberg, C. 1995. Natural vitamin D3 response elements formed by inverted palindromes: polarity-directed ligand sensitivity of vitamin D3 receptor-retinoid X receptor heterodimer-mediated transactivation. Mol.Cell Biol. 15:1154-1161
283. Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human-mouse alignments with BLASTZ. Genome Res. 13:103-107
284. Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. 2000. PipMaker--a web server for aligning two genomic DNA sequences. Genome Res. 10:577-586
285. Sham, M.H., Vesque, C., Nonchev, S., Marshall, H., Frain, M., Gupta, R.D., Whiting, J., Wilkinson, D., Charnay, P., and Krumlauf, R. 1993. The zinc finger gene Krox20 regulates HoxB2 (Hox2.8) during hindbrain segmentation. Cell 72:183-196
286. Sherr, C.J. 1996. Cancer cell cycles. Science. 274:1672-1677
287. Sherr, C.J. and Roberts, J.M. 1999. CDK inhibitors: positive and negative regulators of G1-phase progression. Genes Dev. 13:1501-1512
288. Shibata, R., Misonou, H., Campomanes, C.R., Anderson, A.E., Schrader, L.A., Doliveira, L.C., Carroll, K.I., Sweatt, J.D., Rhodes, K.J., and Trimmer, J.S. 2003. A fundamental role for KChIPs in determining the molecular properties and trafficking of Kv4.2 potassium channels. J.Biol.Chem 278:36445-36454
289. Siddharthan, R. 2006. Sigma: multiple alignment of weakly-conserved non-coding DNA sequence. BMC.Bioinformatics. 7:143

290. Sinclair, A.M., Gottgens, B., Barton, L.M., Stanley, M.L., Pardanaud, L., Klaine, M., Gering, M., Bahn, S., Sanchez, M., Bench, A.J., et al. 1999. Distinct 5' SCL enhancers direct transcription to developing brain, spinal cord, and endothelium: neural expression is mediated by GATA factor binding sites. Dev.Biol. 209:128-142
291. Sinha, S., van Nimwegen, E., and Siggia, E.D. 2003. A probabilistic method to detect regulatory modules. Bioinformatics. 19 Suppl 1:i292-i301
292. Sivak, J.M., West-Mays, J.A., Yee, A., Williams, T., and Fini, M.E. 2004. Transcription Factors Pax6 and AP-2alpha Interact To Coordinate Corneal Epithelial Repair by Controlling Expression of Matrix Metalloproteinase Gelatinase B. Mol.Cell Biol. 24:245-257
293. Smith, T.F. and Waterman, M.S. 1981. Comparison of biosequences. Adv.Appl.Math 2:482-489
294. Smith, T.F. and Waterman, M.S. 1981. Comparison of biosequences. Adv.Appl.Math 2:482-489
295. Song, J., Murakami, H., Tsutsui, H., Ugai, H., Geltinger, C., Murata, T., Matsumura, M., Itakura, K., Kanazawa, I., Sun, K., et al. 1999. Structural organization and expression of the mouse gene for Pur-1, a highly conserved homolog of the human MAZ gene. Eur.J.Biochem. 259:676-683
296. Spina, C.S., Tangpricha, V., Uskokovic, M., Adorinic, L., Maehr, H., and Holick, M.F. 2006. Vitamin D and cancer. Anticancer Res. 26:2515-2524
297. Stahler, P., Beier, M., Gao, X., and Hoheisel, J.D. 2006. Another side of genomics: synthetic biology as a means for the exploitation of whole-genome sequence information. J.Biotechnol. 124:206-212
298. Stevaux, O. and Dyson, N.J. 2002. A revised picture of the E2F transcriptional network and RB function. Curr.Opin.Cell Biol. 14:684-691
299. Stevens, C. and La Thangue, N.B. 2003. E2F and cell cycle control: a double-edged sword. Arch.Biochem Biophys. 412:157-169
300. Stone, J.R. and Wray, G.A. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. Mol.Biol Evol. 18:1764-1770
301. Stormo, G.D. 2000. DNA binding sites: representation and discovery. Bioinformatics. 16:16-23
302. Supplementary Website: http://homes.esat.kuleuven.be/~kmarchal/Supplementary_Information_VanHel_2005/SuppWebsite.html
303. Sze, S.H., Gelfand, M.S., and Pevzner, P.A. 2002. Finding weak motifs in DNA sequences. Pac.Symp.Biocomput.:235-246
304. Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., and Jones, R.T. 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J.Mol.Biol. 203:439-455
305. Taher, L., Rinner, O., Garg, S., Sczyrba, A., and Morgenstern, B. 2004. AGenDA: gene prediction by cross-species sequence comparison. Nucleic Acids Res. 32:W305-W308
306. Tatusova, T.A. and Madden, T.L. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol.Lett. 174:247-250

307. Tavera-Mendoza, L., Wang, T.T., Lallemant, B., Zhang, R., Nagai, Y., Bourdeau, V., Ramirez-Calderon, M., Desbarats, J., Mader, S., and White, J.H. 2006. Convergence of vitamin D and retinoic acid signalling at a common hormone response element. EMBO Rep. 7:180-185
308. Taylor, J.S. and Raes, J. 2004. Duplication and divergence: the evolution of new genes and old ideas. Annu.Rev.Genet. 38:615-643
309. The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 437:69-87
310. The ENCODE (ENCyclopedia Of DNA Elements) Project. 2004. Science 306:636-640
311. The Gene Index Project (TGI): http://compbio.dfci.harvard.edu/tgi/
312. Thijs, G. Probalistic methods to search for regulatory elements in sets of coregulated genes. PhD thesis. faculty of Applied Sciences.Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven. 2003.
313. Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. 2001. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17:1113-1122
314. Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. 2002a. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J.Comput.Biol. 9:447-464
315. Thijs, G., Moreau, Y., De Smet, F., Mathys, J., Lescot, M., Rombauts, S., Rouze, P., De Moor, B., and Marchal, K. 2002b. INCLUSive: integrated clustering, upstream sequence retrieval and motif sampling. Bioinformatics 18:331-332
316. Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788-793
317. Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680
318. Tjian, R. 1996. The biochemistry of transcription in eukaryotes: a paradigm for multisubunit regulatory complexes. Philos.Trans.R.Soc.Lond B Biol Sci. 351:491-499
319. Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. Nat.Biotechnol. 23:137-144
320. Tontonoz, P., Kim, J.B., Graves, R.A., and Spiegelman, B.M. 1993. ADD1: a novel helix-loop-helix transcription factor associated with adipocyte determination and differentiation. Mol.Cell Biol. 13:4753-4759
321. Tran, H., Brunet, A., Grenier, J.M., Datta, S.R., Fornace, A.J., Jr., DiStefano, P.S., Chiang, L.W., and Greenberg, M.E. 2002. DNA repair pathway stimulated by the forkhead transcription factor FOXO3a through the Gadd45 protein. Science. 296:530-534

322. TRANSFAC: http://www.gene-regulation.com/pub/databases.html
323. Travis, A., Amsterdam, A., Belanger, C., and Grosschedl, R. 1991. LEF-1, a gene encoding a lymphoid-specific protein with an HMG domain, regulates T-cell receptor alpha enhancer function [corrected]. Genes Dev. 5:880-894
324. UCR browser: http://mordor.cgb.ki.se/UCRbrowse/
325. UCSC genome browser: http://genome.ucsc.edu/
326. Ureta-Vidal, A., Ettwiller, L., and Birney, E. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat.Rev.Genet. 4:251-262
327. Van de Peer Y., Taylor, J.S., and Meyer, A. 2003. Are all fishes ancient polyploids? J.Struct.Funct.Genomics. 3:65-73
328. Van de Peer, Y., Taylor, J.S., Braasch, I., and Meyer, A. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J.Mol.Evol. 53:436-446
329. van Helden, J., Andre, B., and Collado-Vides, J. 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J.Mol.Biol. 281:827-842
330. Van Hellemont, R., Blomme, T., Van de Peer Y., and Marchal, K. 2007 (in press). Divergence of regulatory sequences in duplicated fish genes. Invited book chapter in 'Genome Dynamics vol. 3: Gene and Protein Evolution'. Editor: Volff, J.-N.
331. Van Hellemont, R., Monsieurs, P., Thijs, G., De Moor, B., Van de Peer Y., and Marchal, K. 2005. A novel approach to identifying regulatory motifs in distantly related genomes. Genome Biol 6:R113.1-R113.18
332. Vandepoele, K., De Vos, W., Taylor, J.S., Meyer, A., and Van de Peer Y. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc.Natl.Acad.Sci.U.S.A 101:1638-1643
333. Venkatesh, B., Gilligan, P., and Brenner, S. 2000. Fugu: a compact vertebrate reference genome. FEBS Lett. 476:3-7
334. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. Science. 291:1304-1351
335. Verlinden, L., Eelen, G., Beullens, I., Van Camp, M., Van Hummelen, P., Engelen, K., Van Hellemont, R., Marchal, K., De Moor, B., Foijer, F., et al. 2005. Characterization of the condensin component Cnap1 and protein kinase Melk as novel E2F target genes down-regulated by 1,25-dihydroxyvitamin D_3. J.Biol Chem. 280:37319-37330
336. Verlinden L, Eelen G, Van Hellemont R, Engelen K, Beullens I, Van Camp M, Marchal K, Mathieu C, Bouillon R and Verstuyf A. 2006. 1alpha,25-Dihydroxyvitamin D(3)-induced down-regulation of the checkpoint proteins, Chk1 and Claspin, is mediated by the pocket proteins p107 and p130. J Steroid Biochem Mol Biol. in press.
337. Verlinden, L., Verstuyf, A., Convents, R., Marcelis, S., Van Camp, M., and Bouillon, R. 1998. Action of 1,25(OH)2D3 on the cell cycle genes, cyclin D1, p21 and p27 in MCF-7 cells. Mol.Cell Endocrinol. 142:57-65
338. Verlinden, L., Verstuyf, A., Van Camp, M., Marcelis, S., Sabbe, K., Zhao, X.Y., De Clercq, P., Vandewalle, M., and Bouillon, R. 2000. Two novel

14-Epi-analogues of 1,25-dihydroxyvitamin D_3 inhibit the growth of human breast cancer cells in vitro and in vivo. Cancer Res. 60:2673-2679

339. Verstuyf, A., Mathieu, C., Verlinden, L., Waer, M., Tan, B.K., and Bouillon, R. 1995. Differentiation induction of human leukemia cells (HL60) by a combination of 1,25-dihydroxyvitamin D_3 and retinoic acid (all trans or 9-cis). J.Steroid Biochem.Mol.Biol. 53:431-441
340. Verstuyf, A., Verlinden, L., Segaert, S., Van Etten, E., Mathieu, C., and Bouillon, R. 1999. Nonclassical effects of 1alpha,25-dihydroxyvitamin D(3) and its analogs. Miner.Electrolyte Metab. 25:345-348
341. Vesque, C., Maconochie, M., Nonchev, S., Ariza-McNaughton, L., Kuroiwa, A., Charnay, P., and Krumlauf, R. 1996. Hoxb-2 transcriptional activation in rhombomeres 3 and 5 requires an evolutionarily conserved cis-acting element in addition to the Krox-20 binding site. EMBO J. 15:5383-5396
342. Vieille-Grosjean, I., Hunt, P., Gulisano, M., Boncinelli, E., and Thorogood, P. 1997. Branchial HOX gene expression and human craniofacial development. Dev.Biol. 183:49-60
343. Villamor, E. 2006. A potential role for vitamin D on HIV infection? Nutr Rev. 64:226-233
344. VISTA genome browser: http://pipeline.lbl.gov/cgi-bin/gateway2
345. Vlieghe, D., Sandelin, A., De Bleser, P.J., Vleminckx, K., Wasserman, W.W., van Roy, F., and Lenhard, B. 2006. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. Nucleic Acids Res. 34:D95-D97
346. Volff, J.N. 2005. Genome evolution and biodiversity in teleost fish. Heredity 94:280-294
347. Voz, M.L., Mathys, J., Hensen, K., Pendeville, H., Van, V., I, Van Huffel, C., Chavez, M., Van Damme, B., De Moor, B., Moreau, Y., et al. 2004. Microarray screening for target genes of the proto-oncogene PLAG1. Oncogene. 23:179-191
348. Vulsteke, V., Beullens, M., Boudrez, A., Keppens, S., Van Eynde, A., Rider, M.H., Stalmans, W., and Bollen, M. 2004. Inhibition of spliceosome assembly by the cell cycle-regulated protein kinase MELK and involvement of splicing factor NIPP1. J.Biol Chem. 279:8642-8647
349. Wadman, I.A., Osada, H., Grutz, G.G., Agulnick, A.D., Westphal, H., Forster, A., and Rabbitts, T.H. 1997. The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. EMBO J. 16:3145-3157
350. Wang, C., Li, Z., Lu, Y., Du, R., Katiyar, S., Yang, J., Fu, M., Leader, J.E., Quong, A., Novikoff, P.M., et al. 2006. Cyclin D1 repression of nuclear respiratory factor 1 integrates nuclear DNA synthesis and mitochondrial function. Proc.Natl.Acad.Sci.U.S.A. 103:11567-11572
351. Wang, Q.M., Jones, J.B., and Studzinski, G.P. 1996. Cyclin-dependent kinase inhibitor p27 as a mediator of the G1-S phase block induced by 1,25-dihydroxyvitamin D3 in HL60 cells. Cancer Res. 56:264-267
352. Washimi, O., Nagatake, M., Osada, H., Ueda, R., Koshikawa, T., Seki, T., Takahashi, T., and Takahashi, T. 1995. In vivo occurrence of p16 (MTS1)

and p15 (MTS2) alterations preferentially in non-small cell lung cancers. Cancer Res. 55:514-517

353. Wasserman, W.W. and Fickett, J.W. 1998. Identification of regulatory regions which confer muscle-specific gene expression. J.Mol.Biol. 278:167-181
354. Wasserman, W.W. and Krivan, W. 2003. In silico identification of metazoan transcriptional regulatory regions. Naturwissenschaften 90:156-166
355. Wasserman, W.W. and Sandelin, A. 2004. Applied bioinformatics for the identification of regulatory elements. Nat.Rev.Genet. 5:276-287
356. Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. 2000. Human-mouse genome comparisons to locate regulatory sites. Nat.Genet. 26:225-228
357. Waterman, M.L. 2002. Expression of lymphoid enhancer factor/T-cell factor proteins in colon cancer. Curr.Opin.Gastroenterol.. 18:53-59
358. Waterman, M.L. 2004. Lymphoid enhancer factor/T cell factor expression in colorectal cancer. Cancer Metastasis Rev. 23:41-52
359. Weinberg, R.A. 1995. The retinoblastoma protein and cell cycle control. Cell. 81:323-330
360. West-Mays, J.A., Zhang, J., Nottoli, T., Hagopian-Donaldson, S., Libby, D., Strissel, K.J., and Williams, T. 1999. AP-2alpha transcription factor is required for early morphogenesis of the lens vesicle. Dev Biol. 206:46-62
361. Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T., and Guigo, R. 2001. SGP-1: prediction and validation of homologous genes based on sequence alignments. Genome Res. 11:1574-1583
362. Wiehe, T., Guigo, R., and Miller, W. 2000. Genome sequence comparisons: hurdles in the fast lane to functional genomics. Brief.Bioinform. 1:381-388
363. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., et al. 2001. The TRANSFAC system on gene expression regulation. Nucleic Acids Res. 29:281-283
364. Wingender, E., Dietze, P., Karas, H., and Knuppel, R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res. 24:238-241
365. Winkler, C., Schafer, M., Duschl, J., Schartl, M., and Volff, J.N. 2003. Functional divergence of two zebrafish midkine growth factors following fish-specific gene duplication. Genome Res 13:1067-1081
366. Wittbrodt, J., Meyer, A., and Schartl, M. 1998. More genes in fish? BioEssays 20:511-515
367. Wittke, A., Weaver, V., Mahon, B.D., August, A., and Cantorna, M.T. 2004. Vitamin D receptor-deficient mice fail to develop experimental allergic asthma. J.Immunol 173:3432-3436
368. Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS.Biol. 3:e7.0116-e7-0130
369. Workman, C.T. and Stormo, G.D. 2000. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Pac.Symp.Biocomput.:467-478

370. Wotton, D., Lo, R.S., Lee, S., and Massague, J. 1999. A Smad transcriptional corepressor. Cell. 97:29-39
371. Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., and Romano, L.A. 2003. The evolution of transcriptional regulation in eukaryotes. Mol.Biol Evol. 20:1377-1419
372. Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434:338-345
373. Yuh, C.H., Bolouri, H., and Davidson, E.H. 1998. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science. 279:1896-1902
374. Zhang, F., Rathod, B., Jones, J.B., Wang, Q.M., Bernhard, E., Godyn, J.J., and Studzinski, G.P. 1996. Increased stringency of the 1,25-dihydroxyvitamin D3-induced G1 to S phase block in polyploid HL60 cells. J.Cell Physiol. 168:18-25
375. Zhang, X., Nicosia, S.V., and Bai, W. 2006. Vitamin D receptor is a novel drug target for ovarian cancer treatment. Curr.Cancer Drug Targets. 6:229-244
376. Zhang, Z. and Gerstein, M. 2003. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J.Biol. 2:11-14
377. Zhang, Z., Gu, J., and Gu, X. 2004. How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? Trends Genet. 20:403-407
378. Zhou, Q. and Wong, W.H. 2004. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc.Natl.Acad.Sci.U.S A. 101:12114-12119
379. Zhu, J. and Zhang, M.Q. 1999. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics. 15:607-611

Curriculum Vitae

Ruth Van Hellemont was born on the 9th of May 1977, in Leuven, Belgium. In 1996, she started her education in applied biological sciences at the K.U.Leuven, where she received the Candidacy diploma in Bioscience Engineering in 1998, and the Masters diploma in Cellular and Biotechnological Engineering in 2001. In 2002, she obtained a Master of Science in Bioinformatics at the K.U.Leuven. Since October 2002 she has been pursuing her PhD as a fellow of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) in the research group ESAT-SCD, under the supervision of Prof. Bart De Moor and Prof. Kathleen Marchal.