Thèse présentée pour obtenir le titre de

Docteur de l'Université Louis Pasteur Strasbourg I

Discipline : Sciences du Vivant
Spécialité : Bioinformatique

par

Aurélie LARDENOIS

Development and applications of an integrated bioinformatics approach for promoter analysis.

Soutenue publiquement le 1er décembre 2006 devant le jury :

Rapporteur externe	Donald J. ZACK
Rapporteur externe	Uwe STRAHLE
Rapporteur interne	Claude KEDINGER
Examinateur	Dino MORAS
Examinateur	Thierry LEVEILLARD
Directeur de thèse	Oliver POCH

à mes parents,

Remerciements

Avant tout, je tiens tout particulièrement à remercier les Professeurs Donald J. Zack, Uwe Strahle, Claude Kedinger, Dino Moras et Thierry Léveillard pour avoir accepté de juger ce travail de thèse.

Je remercie également Dino Moras (doublement donc) pour m'avoir accueilli dans son laboratoire. Merci également à Jean Claude et Annette Thierry pour leur sympathie et pour la qualité irréprochable de leur cuisine bourguignonne.

Je remercie chaleureusement Olivier Poch pour m'avoir donné ce sujet passionnant et m'avoir fait confiance tout au long de cette thèse. Même s'il reste un Himalaya à escalader avant l'excellence, je ne suis déjà plus la même qu'il y a quatre ans et c'est en grande partie grâce à vous. Merci aussi à Claudine pour sa bonne humeur, ses encouragements et ses visites toujours sympathiques.

Merci à tous mes collaborateurs pour cette synergie, pour leur efficacité et leur aide. Tout particulièrement à Nicolas, David, Pascal et Emmanuel de l'équipe de Jean-Marc Egly, à Sophie et Sacha de l'équipe de Thierry Léveillard, à l'équipe de Daniel Boujard, à Michael Primig, à Valeria Roni, à Katia et Jérôme de l'équipe de Brigitte Kieffer, à Gretta et Yvon Trottier.

Gretta, c'était un vrai plaisir de travailler avec toi. Je te souhaite bonne chance pour la suite et espère te revoir bientôt au Liban.

Valéria, à quand une nouvelle visite de Big Apple et un congrès entre filles !

Merci à toute l'équipe de bioinfo au sein de laquelle il fait si bon travailler !

Un grand merci à Raymond, l'homme aux mille et une facettes, programmeur le jour et réparateur de citröen AX la nuit. Isabelle, François et Sophie, merci également pour votre sympathie, pour nous avoir fait découvrir les cours de rock et pour nos soirées chaleureuses en chanson jusqu'à pas d'heure !

Merci M'sieur Luc pour toutes nos conversations de travail et ton côté fraternel toujours appréciable dans les moments de doute. J'oubliais, merci également pour ces beuglantes

soirées chez toi au grand désarroi de tes voisins. Ah, au fait, il n'y a pas de doute le roi Louis….c'est toi !

Merci Julie pour ton aide au cours de ma thèse. Mille mercis pour toutes les corrections d'anglais et l'investissement que tu as fait sur ce manuscrit, avec le sourire, sans jamais te plaindre et en trouvant toujours les bons mots pour me soutenir !

Merci Odile pour ta gentillesse et pour m'avoir réconforté à chaque fois que j'en ai eu besoin au cours de cette thèse. Beaucoup de bonheur à la petite nouvelle du labo : Maëlle !

Merci à Yann pour les bons moments passés au cours de rock et de lindy ! Bon courage pour ta thèse, mais fais attention à ne pas passer toutes tes nuits au labo…

Merci Véro pour ta joie de vivre et ta pêche naturelle. J'ai été vraiment heureuse de pouvoir montrer au gars du labo qu'en ping-pong nous n'étions pas des fillettes ;-) (comme eux !).

Merci Annaïck pour ta gentillesse, ton soutien et nos discussions pendant les pauses thé.

Merci Anne pour ta bonne humeur et courage pour tous les articles et bouquins que tu rédiges en anglais bien sûr et aussi pour la suite de ta thèse!

Merci Nicolas pour ton goût invétéré pour les films exotiques qui contribue à la bonne humeur du labo.

Merci aux travailleurs du couloir infernal pour leur sympathie et leur bonne humeur : Guillaume, Manu et Laetitia.

Fred et Laurent, merci pour nos discussions au RU et votre aide précieuse pour les installations de banques, de programmes et autres analyses de promoteurs.

Merci à mon premier voisin de bureau, à savoir le maître Jedi Dr Finton Sirockin pour sa sympathie et son humour.

Merci Wolfgang pour nos discussions enrichissantes et d'avoir partagé ce bureau.

Ravi, bonne chance pour cette dernière année de thèse !

Merci à Christiane, Anne, Laetitia et Odile pour votre gentillesse et votre efficacité.

Merci à Serge et Guillaume pour votre aide précieuse.

Courage aussi à ceux qui reprendront le flambeau de la promotologie ;-). N'oubliez surtout jamais d'avoir un aimant sur vous afin de retrouver l'aiguille…

Merci à Wiwi (BP) et Murielle pour votre sympathie, vos encouragements et vos bon plans restos et autres (je n'inclus pas les blagues téléphoniques ;-))

Merci à Anne et Rachelle (et Céline) pour nos folles soirées « poule », tricot et compagnie… Merci pour votre gentillesse, votre bonne humeur et votre soutien ! Et maintenant célébrons l'arrivée d'une toute nouvelle poulette Anna !

Diane et Emeric (les lapins Joe Barr et leur panthère), Mélinda et Thierry, Vivi et Olive, tous bien loin du grand est et pourtant toujours aussi encourageants. Merci pour votre amitié…

Jean et Céline, par quoi commencer… Merci pour nos repas « Made in Strasbourg », nos soirées de jeux endiablées, nos franches rigolades (Sans rancune Jean si c'est souvent toi qui trinque ;-)). Merci surtout pour votre amitié et comme dirait l'autre « Pourvu que ça dure ! ». Enfin, bienvenue à Anna et beaucoup de bonheur à vous trois ! euuuuh…, que dis-je …à vous quatre, avec Sully bien sûr !

Merci à ma « jolie famille » : Marinette, Yvounet, Toto, Pierrot la Lune et Nanou pour votre soutien depuis toutes ces années.

Un énorme merci à mes parents pour tout, votre soutien sans faille depuis toujours. Merci d'avoir accepté, soutenu sans réserve mes orientations et de m'avoir permis d'en arriver là,…Je vous en suis très reconnaissante et vous aime profondément.

Merci à mon frère adoré, Gaston pour ton soutien depuis toutes ses années, pour m'avoir donné l'envie d'aller plus loin et de marcher sur tes traces. Merci à toi et à Agnès pour vos encouragements.

Un grand merci à mes grands-parents, les moments passés auprès de vous sont rares mais tellement importants !

Merci à mon cœur… mon amour ! Mon amour…, mon cœur !... mon clown, mon professeur, mon tout. Ton soutien et ton aide ont été primordiaux pour ce travail. Tout au long de cette thèse, tu as toujours su me redonner le moral et m'encourager. A nous maintenant d'en profiter et de renouveler les moments merveilleux que l'on a passé à l'autre bout de la terre…

Enfin, merci à monsieur Royco qui m'a accompagné tout au long de cette thèse ainsi qu'à Monsieur Picard et Mlle C. Vitamine qui nous ont permis de survivre ces derniers mois.

Mes pensées vont aussi à Jacqueline et Jacques qui nous ont quittés beaucoup trop vite au cours de cette thèse et nous manquent profondément.

"Nothing makes sense in biology except in the light of evolution"

Theodosius Dobzhansky's dictum

List of abbreviations

5′ / 3′ UTR: 5′ / 3′ UnTranslated Region

bp: base pair

BRE: TFIIB Recognition Element

CAGE: Cap Analysis of Gene Expression

CDS: Coding sequence

CNS: Conserved Noncoding Sequences

CRM: *Cis*-Regulatory Modules

DBD: DNA-Binding Domain

DNA: DesoxyriboNucleic Acid

DPE: Downstream Promoter Element

EST: Expressed Sequence Tags

Go: Giga Octets

GUI: Graphical User Interface

INR: initiator element

kb: kilo base pair

LBD: Ligand-Binding Domain

(m)RNA: (messanger) RiboNucleic Acid

miRNA: MicroRNA

MACS: Multiple Alignments of Complete Sequences

MTE: Motif Ten Element

NCR: Non Coding Region

NR: Nuclear Receptor

ncRNA: non-protein coding RNA

PFM: Position Frequency Matrix

PIC: PreInitiation Complex

PPR: Putative Promoter Region

PSSM: Position Specific Scoring Matrix

PWM: Position Weight Matrix

RE: Regulatory Element

SAGE: Serial Analysis of Gene Expression

S/MARs: Scaffold/Matrix Attachment Regions

SINES: short interspersed elements

TF: Transcription Factor

TFBS: Transcription Factor Binding Site

TSS: Transcriptional Start Site

Table of contents

Table of figures

Table of tables

Résumé

Depuis la mise en évidence de l'ADN comme source première de l'information génétique et la détermination, en 1953, de la structure de la double hélice d'ADN, la bioinformatique est devenue une discipline à part entière dans la recherche et les développements des sciences du vivant. Initialement conçue autour de méthodes informatiques dédiées à l'organisation et à l'analyse des données déposées dans les premières bases de données biologiques, la bioinformatique s'est structurée, dans le courant des années 80, autour de différents champs d'application pour aboutir à une discipline de recherche indépendante. Schématiquement, trois branches majeures sont souvent distinguées correspondant aux aspects de stockage et de récupération des données, aux aspects de traitements et d'analyses statistiques et informatiques des données et enfin, ceux couvrant le développement de nouveaux algorithmes de prédiction à même de fournir de nouvelles informations. A l'ère post-génomique, la bioinformatique est traversée par une véritable révolution liée à la disponibilité de nombreuses séquences de génomes complets coïncidant avec la production d'une vaste quantité de données liées à l'émergence de technologies à haut débit. Chez les eucaryotes supérieurs, ces développements permettent d'envisager pour la première fois, des études de systèmes biologiques complexes en ouvrant la perspective d'une compréhension fine des réseaux de régulation de l'expression des gènes.

Dans ce cadre, l'identification de promoteurs potentiels et la caractérisation des sites de fixation de facteurs de transcription (Transcription Factor Binding Sites, TFBSs) est un challenge de première importance. Ces recherches constituent actuellement un enjeu considérable dont une des finalités est une meilleure appréhension des voies et processus biologiques et la découverte de nouvelles approches thérapeutiques contrôlant l'expression des gènes.

La bioinformatique dédiée à l'analyse in silico de promoteurs constitue un domaine en pleine effervescence ces dernières années. De nombreux outils spécifiquement dédiés à cette thématique ont ainsi été développés afin de prédire des promoteurs et des TFBSs potentiels. Néanmoins, les TFBSs étant de courtes séquences (5 à 20 paires de bases) hautement dégénérées, ces outils prédisent fréquemment plus de 90 % de faux positifs. Face à cette quantité importante de motifs potentiels mais non biologiquement actifs, il est primordial d'élaborer des stratégies permettant d'augmenter le rapport signal/bruit.

La méthode dite du «phylogénétique footprinting» est communément admise dans le domaine pour remédier à ce problème et ainsi améliorer la prédiction de TFBSs potentiellement actifs. Cette méthode, basée sur la réintégration des promoteurs dans leur

contexte évolutif, émet l'hypothèse que les sites de régulation présents dans les promoteurs sont soumis à une pression de sélection au cours de l'évolution bien supérieure à celle appliquée aux éléments non fonctionnels. L'alignement de promoteurs de divers organismes permet ainsi de délimiter les zones promotrices potentiellement actives et d'y identifier des TFBSs potentiellement actifs en fonction de leur conservation au cours de l'évolution.

Dans le cadre de ma thèse, face à l'émergence de nombreux outils et de données dédiés à l'analyse des promoteurs, il nous est apparu essentiel d'envisager une logique intégrative combinant et évaluant l'ensemble des informations expérimentales et prédites. Cette conception de l'analyse des promoteurs a donné lieu au développement de PromAn (http://bips.u-strasbg.fr/PromAn), programme modulaire qui permet la juxtaposition et l'intégration de différentes étapes d'analyse.

Brièvement, PromAn valide tout d'abord la localisation du site d'initiation de la transcription (Transcription Start Site, TSS) par l'utilisation de divers programmes de prédiction mais aussi par des recherches dans des banques de données expérimentales. Comme la localisation d'un promoteur est directement fonction de la position du TSS, cette étape de validation précise du TSS est primordiale dans la mesure où de nombreuses séquences de transcripts (ARNm) présentes dans les banques de données ne sont pas complètes en 5'. En effet, ces régions manquantes des ARNm peuvent en réalité correspondre à un ou plusieurs exons non-codants supplémentaires décalant ainsi la position réelle du TSS, et donc du promoteur, de plusieurs milliers de paires de bases.

Une fois la localisation des promoteurs achevée, la méthode de «phylogénétique footprinting» est appliquée afin de localiser les zones promotrices conservées au cours de l'évolution et donc potentiellement fonctionnelles. De plus, les motifs des TFBSs de diverses banques de données et extraits de la littérature, sont prédits sur le promoteur étudié et évalués en fonction de leur conservation au cours de l'évolution. PromAn est disponible par le biais d'un serveur web (http://bips.u-strasbg.fr/PromAn) dédié à l'analyse d'un gène isolé. Une version locale, impliquant moins de contraintes que la version web et intégrant des programmes supplémentaires, a également été développée. PromAn peut ainsi être utilisé pour réaliser des analyses à haut débit. Une interface graphique intuitive permet à l'utilisateur de visualiser, d'interpréter et de sélectionner les résultats selon divers paramètres.

La philosophie qui a sous-tendu les analyses bioinformatiques de promoteurs effectuées au moyen de PromAn a consisté à coupler systématiquement prédictions et validations expérimentales afin de corroborer ou non les prédictions obtenues. Ce processus d'analyse en va et vient entre prédiction et expérimentation est non seulement une étape essentielle

pour la caractérisation précise et complète d'un système de régulation donné mais également, une stratégie optimale pour améliorer et optimiser la création d'un programme bioinformatique. Dans ce cadre, au travers de diverses collaborations avec des équipes de biologistes, PromAn a été mis à profit afin d'identifier, d'analyser et/ou de caractériser les promoteurs potentiels de différents types de gènes cibles. En effet, ces études ont porté sur différents systèmes reflétant différents niveaux de complexité dans le domaine de la régulation des gènes. Ainsi, dans un premier temps, nos analyses se sont concentrées sur un petit groupe de gènes (FABP4 et HMGCS2). Ces gènes ont été traités par PromAn dans un cadre bien défini de comparaison simple, homme/souris, suivi de la recherche et de l'affinement des régions génomiques « enhancers » conservées au cours de l'évolution. Ces prédictions ont été confortées à des expériences de validation in vitro qui ont permis de vérifier la puissance et la fiabilité des résultats fournis par PromAn.

Dans une seconde série d'études, des systèmes plus complexes ont été abordés englobant aussi bien des familles de gènes proches que des groupes de plusieurs milliers de gènes co-exprimés provenant d'analyses de données de transcriptomique. Ces approches se sont focalisées sur l'étude d'un même système biologique complexe, celui de la vision, thématique majeure de notre laboratoire et notamment sur les gènes de la vision que l'on retrouve dérégulés dans diverses maladies, telles que la rétinite pigmentaire ou les maladies liées à la présence d'expansion de polyglutamine dans des gènes mutés. L'analyse d'un grand nombre de gènes a permis d'aborder la problématique de la validation statistique des informations biologiques à haut débit. Ceci a abouti au développement de différents outils d'estimation du rapport signal/bruit, dont certains ont été à la base des modules de validation statistique de PromAn.

La rétinite pigmentaire, maladie monogénique, affecte plus de 40000 patients en France et son évolution graduelle mène à la perte totale de la vision en deux étapes successives. Tout d'abord la perte des bâtonnets, responsables de la vision périphérique, conduit à la réduction du champ visuel et à une vision dite «tubulaire» basée sur la seule activité des cônes, responsables de la vision centrale et nocturne et de la détection des couleurs. Enfin, la perte des cônes conduit à la cécité complète. Un facteur trophique, RdCVF (Rod-derived Cone Viability Factor), sécrété par les bâtonnets et assurant la survie des cônes a été identifié. Dans ce contexte, j'ai participé à une étude bioinformatique approfondie de cette protéine qui a amené à l'identification d'un paralogue très prometteur (RdCVF2) qui possède les mêmes propriétés fonctionnelles ainsi que les mêmes structures génique et protéique. Les facteurs trophiques RdCVFs constituent des cibles thérapeutiques privilégiées et la connaissance des mécanismes de régulation de l'expression de ces deux gènes est primordiale pour envisager

des voies thérapeutiques permettant de maintenir la vision tubulaire. Là encore, PromAn nous a permis de mettre en évidence des promoteurs conservés au cours de l'évolution et surtout, de révéler la présence de TFBSs potentiellement fonctionnels et responsables de l'expression de chacun de ces gènes très proches. Les validations expérimentales en cours ont permis de vérifier que des signaux de régulation semblaient présents dans les promoteurs humains, cependant que les promoteurs murins paraissent plus difficiles à caractériser. Devant ces résultats prometteurs, plusieurs séries de validation expérimentale sont en cours afin d'affiner les interprétations et de mieux comprendre le système de régulation RdCVF/RdCVF2 dans la rétine et son rôle dans le rapport de dépendance trophique cône/bâtonnet.

L'augmentation du rapport signal/bruit dans l'étude de voies de régulation peut également être envisagée par la recherche de TFBSs communs à des groupes de gènes co-exprimés impliqués dans les même processus biologiques. Des estimateurs statistiques ont été utilisés pour caractériser fonctionnellement des groupes de gènes co-exprimés par l'utilisation de banques d'annotation spécialisées telles que la Gene Ontology. La Gene Ontology correspond à un vocabulaire standardisé et hiérarchisé définissant les protéines/gènes selon trois catégories : la localisation cellulaire, la fonction moléculaire et le processus biologique. Afin d'améliorer la qualité et la quantité de ces annotations pour chaque protéine/gène identifié(e), j'ai été amenée à participer au développement d'un nouvel algorithme de propagation de l'annotation, GOAnno (http://bips.u-strasbg.fr/GOAnno). Ce programme est notamment basé sur l'utilisation d'alignements multiples de séquences complètes de qualité et hiérarchisés en sous-famille, une des thématiques centrales de notre laboratoire (http://bips.u-strasbg.fr/PipeAlign). Une fois cette étape achevée sur l'ensemble des gènes co-exprimés, une caractérisation fonctionnelle de chaque groupe permet de faciliter la compréhension du système biologique étudié. Une annotation, par exemple un terme ontologique, est représentative d'un groupe de gènes présentant des profils d'expression similaires dans la mesure où elle y est enrichie. Cet enrichissement est évalué par des estimateurs statistiques tels que le z score ou la p-value. Cette démarche a notamment été utilisée dans le cadre d'une étude de transcriptomique visant à comprendre la toxicité induite par des expansions de polyglutamine impliquées dans la maladie de Huntington ou l'ataxie spinocérébelleuse de type 7. Cette analyse a mis en évidence différents groupes de gènes co-régulés présentant des enrichissements en divers processus biologiques qui ont notamment permis de démontrer que l'expansion polyglutamine est responsable d'une perte progressive de l'expression de gènes spécifiques des bâtonnets matures en compromettant le programme génétique impliqué dans la maintenance de l'état différencié de ces

photorécepteurs. Ces mécanismes ont pour conséquence le disfonctionnement suivi de la dégénérescence des bâtonnets.

Cette étude à haut débit a clairement démontré une relation étroite entre co-expression et processus biologique illustrant la qualité des données ainsi que la robustesse et la stringence de la définition des groupes de co-expression. Nous avons dès lors testé l'hypothèse que ces profils d'expression similaires pourraient être le fruit de voies de régulation sous la dépendance de facteurs de transcription communs. Dans un premier temps, nos efforts ont portés sur l'étude du facteur de transcription stat3 surexprimé de façon aberrante dans l'analyse transcriptomique réalisée et connu pour être impliqué dans des mécanismes d'inhibition de la différenciation des photorécepteurs. Pour réaliser cette analyse, la version à haut débit de PromAn a été mise à profit en intégrant des adaptations des estimateurs statistiques précédemment décrits. Cependant cette étude, qui visait à mettre en évidence des enrichissements éventuels de gènes possédant le motif de régulation stat3 dans les différentes populations de gènes co-exprimés, n'a pas permis d'établir de façon non ambiguë une telle sur-représentation dans aucun groupe de gènes.

Ces résultats négatifs nous ont incités à entreprendre d'autres séries d'analyse de promoteurs à haut débit, afin d'établir si l'absence de mise en évidence d'une sur-représentation des signaux de régulation dans les groupes de gènes était liée au système biologique complexe étudié ou à des limitations de la version haut débit du programme PromAn. Ces études ont abordées deux systèmes biologiques distincts : la rétine et la lignée germinale chez le mâle. Dans chaque cas, les données de transcriptomique disponibles impliquaient un grand nombre de conditions expérimentales (temporelles ou tissulaires) et les analyses aboutissaient à des groupes de gènes co-régulés pour lesquels des enrichissements en processus biologiques étaient statistiquement significatifs. Ces groupes de gènes co-régulés ont dès lors été étudiés par la méthode à haut débit de PromAn. Ceci a permis de mettre en évidence dans chaque cas, des enrichissements statistiquement significatifs en promoteurs présentant des motifs de facteurs de transcription directement impliqués dans le développement de la rétine (Crx) ou dans la spermatogenèse (CREM).

L'ensemble de ces résultats illustre l'efficacité de l'approche intégrative du programme PromAn non seulement lors d'analyses approfondies de promoteurs de gènes-cibles mais aussi lors d'études à haut débit s'appuyant sur des groupes de gènes co-régulés issus d'analyses des données de transcriptomique. De plus, la robustesse de cette approche permet d'envisager des études plus ambitieuses visant à caractériser des réseaux de régulation impliqués dans de grands processus biologiques, tels que le développement d'un tissu ou le développement des maladies humaines.

Introduction

Chapter 1 - Biology and bioinformatics context

In the 50s, biology and our understanding of life were deeply affected by two major findings:

- the discovery of DNA (DesoxyriboNucleic Acid) as the source of genetic information together with the elucidation of the double-helical nature of the DNA molecule (Watson and Crick, 1953).
- the central dogma of molecular biology, first enunciated by Francis Crick in 1958 (Crick, 1958). This dogma emphasized the central role of genes and proteins in cellular organization (Figure 1).

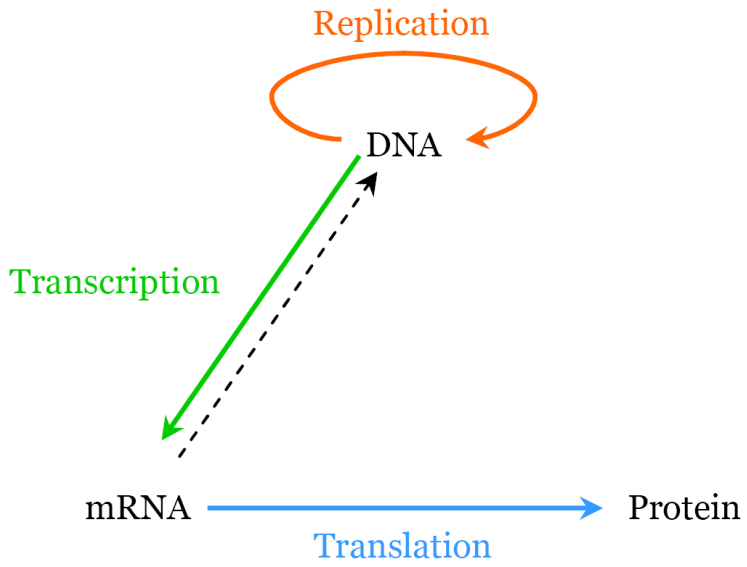


Figure 1. The central dogma of molecular biology: flow of genetic information.

This dogma forms the backbone of molecular biology by explaining the flow of genetic information from DNA to RNA to protein, and includes several major processes:

- Replication: DNA replicates its information in a process that involves many enzymes.
- Transcription: DNA codes for the production of messenger RNA (mRNA). In eucaryotic cells, mRNA is processed (essentially by splicing) and migrates from the nucleus to the cytoplasm.
- Translation, ribosomes "read" the information carried by the mRNA and use it for protein synthesis.

Proteins are involved in almost all biological activities, structural or enzymatic.

This revolutionar concept of the flow of genetic information, together with reverse transcription discovered later on, coincided with a rapid evolution in the research and development of novel techniques not only in biology and informatics but also in a new scientific domain: bioinformatics which has since become an integral part of biological sciences (Table 1).

When	Who	What
1953	Watson	Double helix model for DNA (Watson and Crick, 1953)
	IBM 650	First commercial computer
	Sanger	Determination of the amino acid sequence of the A and B chains of insulin (Sanger and Thompson, 1953)
1956	Anfinsen	Three-dimensional conformation of proteins is specified by their amino acid sequence (Anfinsen and Redfield, 1956)
1958	Crick	Enunciated the central dogma of molecular genetics: information flows from DNA to RNA to protein (Crick, 1958)
1961-1965	Nirenberg and Matthaei	Genetic code words for the amino acids identification (Matthaei et al., 1962)
1965	Dayhoff	The first Atlas of Protein Sequence and Structure, which contained sequence information on 65 proteins (Dayhoff, 1965)
1967	Fitch	Calculation of phylogenetic relationships of twenty organisms, ranging from fungi to mammals, by comparing their cytochrome C amino acid sequences (Fitch and Margoliash, 1967)
1969	ARPANET	Computers linking between several universities
1970	Needleman and Wunsch	Algorithm for global optimal alignment between two protein sequences (Needleman and Wunsch, 1970)
1974	Cerf and Kahn	Development of the concept of connecting networks of computers into an "internet" and develop the Transmission Control Protocol (TCP)
	Chou	Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins (Chou and Fasman, 1974)
1977	Sanger	Rapidly sequencing methods of long nucleic sections (Sanger et al., 1977)
1978	Sanger	The first complete genome sequence for virus (pi-x 174) (5386 base pairs, bp) (Sanger et al., 1978)
1980	EMBL	Creation of the first European nucleic sequences database
1981	Smith and Waterman	Algorithm for local optimal sequence alignment (Smith and Waterman, 1981)
	IBM	IBM introduces its Personal Computer to the market
1982	Genbank	Creation of the American database of nucleic sequences
1983	Mullis	Invention of the polymerase chain reaction (PCR)
1984	Gouy	ACNUC: software of inquiry of sequences databases (Gouy et al., 1984)
1985	Lipman and pearson	FASTA (Lipman and Pearson, 1985)
1986	SWISS-PROT	Creation of the protein sequences database
	DDBJ	Creation of the Japanese nucleic sequences database
	Thomas Roderick	Notion of the “genomic” term
1987	Applied Biosystems	Marketing of the first automated sequencer
	Burke	Creation of the cloning vector YAC (Yeast Artifical Chromosome) (Burke et al., 1987)
	McKusick	First genetic map of the human genome (McKusick and Ruddle, 1987)
	DNA chip	Creation of the technology

1988	HUGO	Coordinates the mondial sequencing of the human genome
	Higgins	CLUSTAL: software for multiple alignment of sequences (Higgins and Sharp, 1988)
	Peterson	Use of the Taq polymerase for the PCR (Peterson, 1988)
	Wingender	TRANSFAC database (Wingender, 1988)
	Tagle	Phylogenetic footprinting (Tagle et al., 1988)
1989	Fields	Double-hybrid system allowing the detection of interaction between two proteins
	Internet	Internet replaces ARPANET
1990	Altschul	Blast (Altschul et al., 1990)
	Berners-Lee	First HTML document publication
	Collins	Positional cloning development (Collins, 1990)
	HGP	"Human Genome Project" aiming at sequencing the whole human genome
1991	Adams	First high-throughput sequencing of cDNA (EST) (Adams et al., 1991)
	Roberts	GRAIL: gene localization program
1992		Sequencing of the chromosome III of *Saccharomyces cerevisiae*
1993	Cohen	Firt physical map of the human genome (Cohen et al., 1993)
	Boguski	dbEST: international database of EST (Boguski et al., 1993)
	Etzold	SRS: program for databases request (Etzold and Argos, 1993)
1994	Thompson	ClustalW, multiple sequence alignment (Thompson et al., 1994)
1995	Fleischmann	Sequencing of the first living organism, the bacteria *Haemophilis influenzae* (1,8 Mbp) (Fleischmann et al., 1995)
1996	Walsh	Sequencing of the first eukaryotic genome *Saccharomyces cerevisiae* (12.1 Mbp) (Walsh and Barrell, 1996)
	Affymetrix	Marketing of the first DNA chip
1997	Altschul	Gapped Blast (Altschul et al., 1997)
	Blattner	*Escherichia coli* sequencing (4.7 Mbp) (Blattner et al., 1997)
	Burge	GenScan program program: complete gene structures prediction in human genomic DNA (Burge and Karlin, 1997)
1998		Sequencing of the first pluricellular organism, *Caenorhabditis elegans* (97 Mbp)
1999	Dunham	Sequencing of the human 22 chromosome (Dunham et al., 1999)
2000	Stover	*Pseudomonas aeruginosa* genome sequencing (6.3 Mbp) (Stover et al., 2000)
	Dennis	*Arabidopsis thaliana* genome sequencing (100 Mbp) (Dennis and Surridge, 2000)
	Adams	*Drosophila melanogaster* genome sequencing (180 Mbp) (Adams et al., 2000)
2001	Lander	Preliminary sequence of the human genome (3 Gbp) by HGP (Lander et al., 2001)
	Venter	Preliminary sequence of the human genome (3 Gbp) by Celera Genomics (Venter et al., 2001)
2002	Waterston	Preliminary sequence of the mouse genome (2.5 Gbp) (Waterston et al., 2002)
	Suzuki	Database of Transcriptional Start Sites, DBTSS (Suzuki et al., 2002)
2003	Boffelli	Phylogenetic shadowing method (Boffelli et al., 2003)
	Schwartz	BLASTZ pairwise alignment tool (Schwartz et al., 2003)
	Lenhard	CONSITE, phylogenetic footprinting tool (Lenhard et al., 2003)
	Karolchik	UCSC Genome Browser database (Karolchik et al., 2003)
2004	Gibbs	Sequence of the whole rat genome (Gibbs et al., 2004)
	Jaillon	Sequence of the whole genome of *Tetraodon nigroviridis* (385 Mbp) (Jaillon et al., 2004)
	Hillier	Sequence of the whole genome of the chicken, *Gallus gallus* (Hillier et al., 2004)

	Blanchette	MULTIZ, TBA, local alignment programs (Blanchette et al., 2004)
	Stalker	ENSEMBL Genome Browser (Stalker et al., 2004)
Soon…		Further mammalian complete genomes oh higher eukaryotes sequenced as several primates, the pig, the cow, the horse, the kangaroo, the elephant, the sheep, the dog, the cat, the rabbit, the frog, the zebrafish etc (more than 600 eukaryote complete genomes in current sequencing)

Table 1. Non exhaustive chronology of critical events in biology, informatic and bioinformatics. The writing colors black, green and blue describe events in biology, informatic and bioinformatics respectively. Additionally, advances in promotology are depicted in orange. They are described in the following parts of this manuscript. This table is adapted from http://bio.cc/Bioinformatics/history_of_bioinformatics.html .

Initially, computational methods were developed to organise and analyse the data stored in the first biological databases and, in the 80's, the field of bioinformatics took shape as an independent research discipline. For the first time, efficient algorithms were developed to cope with an increasing volume of information, and their computer implementations were made available for the wider scientific community. Briefly, three major problems were addressed at this time: the storage and retrieval of the data, the computational and statistical analyses and the algorithms for prediction.

In 1977, two high-throughput sequencing methods were simultaneously developed: an enzymatic approach (Sanger et al., 1977) and a chemical approach (Maxam and Gilbert, 1977). As this last method is toxic for the research worker, the former became more widely used and gave the biologists the opportunity to sequence their genes of interest. Nevertheless, the sequencing of complete genomes remained a difficult challenge. Excluding small viral genomes, the first complete genome sequence of an organism belonging to one of the three domains of life (*Bacteria, Archaea and Eukarya*) (Figure 2) (Woese et al., 1990) was the bacterial *Haemophilus influenzae* genome (Fleischmann et al., 1995).

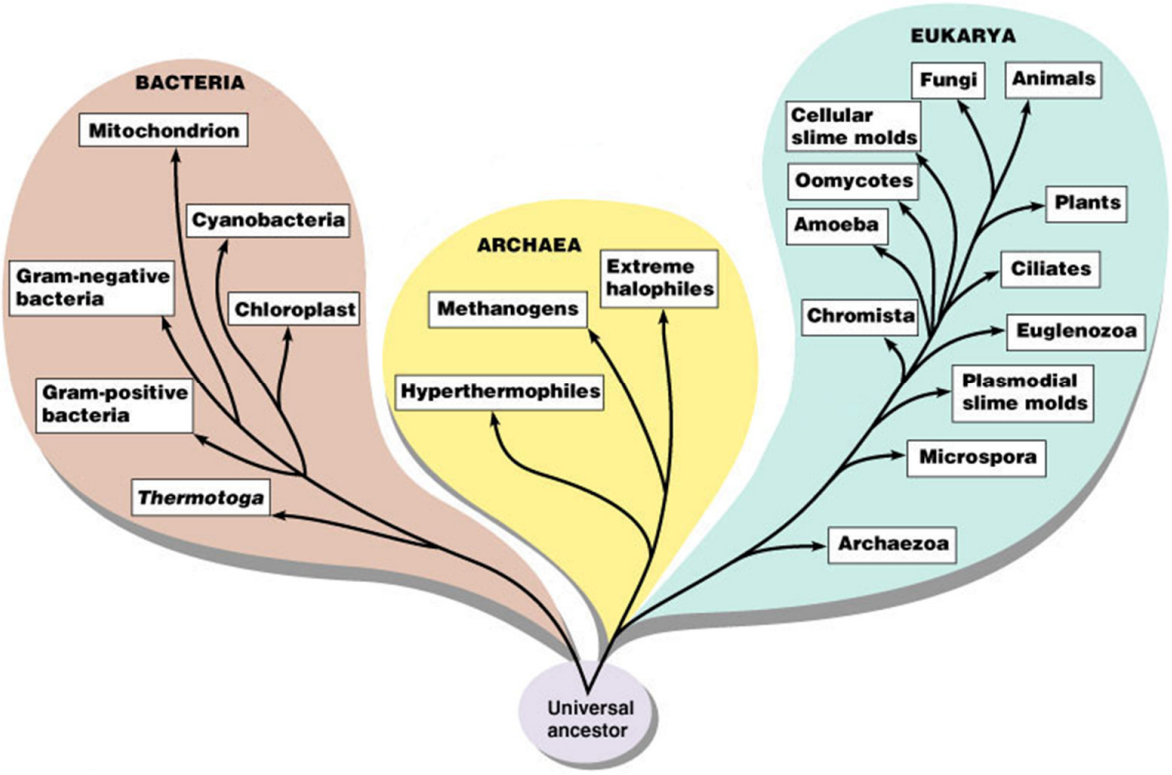


Figure 2. Tree of life.
This schema represents a simplified version of the universal tree of life. The three domains of life, Bacteria, Archaea and Eukarya are color-coded in salmon, yellow and cyan respectively.

Since then, the sequencing of prokaryotic and simple eukaryotic genomes rapidly increased although the complete human genome with its 3 billion base pairs remained elusive. The Human Genome Project (HGP), initiated at the Imperial Cancer Research (ICR) in London and coordinated by HUGO (Human Genome Organization), together with the first automated sequencers in 1987, resulted in the publication of the first human genome draft sequence in 2001 (Lander et al., 2001) (Venter et al., 2001). This huge work now provides a unique opportunity to study the complete set of human genes, their function and regulation in their genomic context. However, this version of the human genome was a draft, and much work still remains to be done before the obtention of a complete sequence of a high quality human genome with no gap or ambiguities.

The end of the 20th century was characterized by the emergence of large-scale EST (Expressed Sequence Tags) and complete genome sequencing projects (Figure 3).

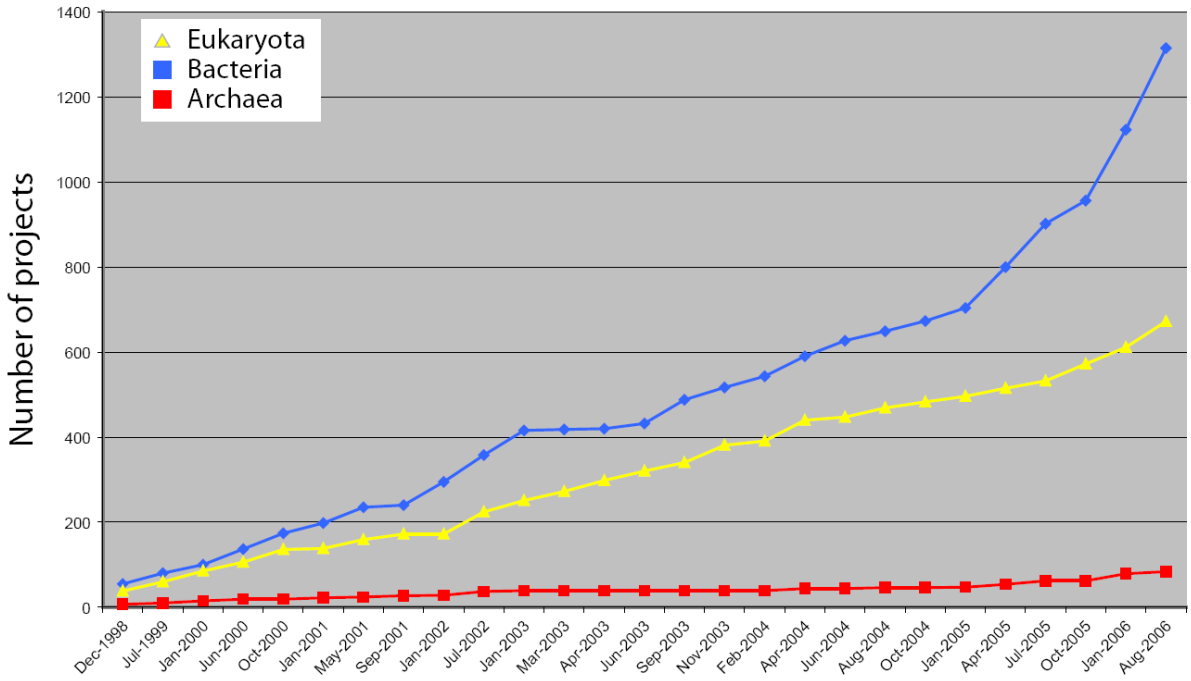


Figure 3. Evolution of the number of sequencing projects of complete genomes as well as of Expressed Sequence Tags (EST) in the three domains of life (data available on the GOLD web server).

Thanks to the technical progress of large-scale sequencing, genome and EST sequencing projects are experiencing an exponential increase, especially in the *Bacteria* and the *Eukaryota* domains of life.

As a consequence, sequences of 405 eukaryotic genomes are now available to the scientific community (GOLD web server at the adress http://www.genomesonline.org, June 2006), including 27 *Archaea*, 337 *Bacteria* and 31 *Eukaryota*. In the *Eukaryotic* domain, several vertebrate genomes are now being sequenced, including several primates: dog, cow, rabbit, pig, elephant, opossum and chicken.

Information concerning the completely sequenced genomes and the ongoing projects is available from several web servers of sequencing centers or dedicated to given genomes (Table 2).

Web server name	Web server address
GOLD	http://www.genomesonline.org
NCBI	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
EBI	http://www.ebi.ac.uk/genomes
TIGR	http://www.tigr.org/tdb
SANGER	http://www.sanger.ac.uk/Projects
JGI	http://www.jgi.doe.gov

Table 2. Non exhaustive list of web servers referencing the list of sequenced genomes and of genomes in sequencing.

Bioinformatics has been profoundly transformed by the post-genomic era which relies on the availability of numerous complete genome sequences, as well as on the new information produced by the emerging high-throughput technologies (transcriptomics, proteomics, interactomics, etc). As a consequence, the biological databases are now submerged with a mixture of experimental data and computational predictions or analyses with their inherent unreliability. In this context, information management systems are being introduced to collect and store as well as to curate and cross-validate all this heterogeneous information allowing efficient data retrieval and exploitation. These developments are opening up the possibility of new large scale studies of complex biological systems.

From a scientific point of view, many major insights have resulted from the availability of whole genome sequences, for example, the discovery that gene structure in higher eukaryotes is extremely variable in their size, in their number of exons and introns, in their organization, etc. To illustrate this, some statistics are presented for the human genome build 36.1 (Table 3).

	Human genome: 21571 known coding genes			
	Count	Min	Max	Average
Exons	-	2 bp	20,228 bp	288 bp
Introns	-	2 bp	721,292 bp	5,656 bp
One-exon genes	3,068	-	-	-
One-exon transcripts	3,094	-	-	-
Exons per gene	-	0	316	7

Table 3. Statistics on the human genome.
The number of known coding genes was extracted from the Ensembl Genome Browser at http://www.ensembl.org/Homo_sapiens/index.html. These statistics are available at the NCBI (National Center for Biotechnology Information) web servers http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=36&ver=1

As an example of diversity, the smallest protein-coding gene in the human genome is only 500 nucleotides long and has no intron. It encodes a histone protein. In contrast, the largest human gene is 2.5 million nucleotides in length and encodes the dystrophin protein (40

exons). This protein, when missing or non-functional, is responsible for human muscular dystrophy diseases.

The availability of complete genome sequences also provides the opportunity to perform comparative multiple sequence analysis at a genome level (Hardison et al., 2003). During evolution, various large-scale processes, such as recombination, deletion or horizontal transfer, took place causing frequent genome rearrangements and modifications (Shapiro, 2005). Comparative analyses of complete genomes present a comprehensive view of the conservation of gene order, or synteny, between different genomes, and thus provide a measure of organism relatedness at the genome scale (Elnitski et al., 2005) (Ye and Huang, 2005). Examples of such genome level analyses include comparisons among enteric bacteria (McClelland et al., 2000) or between mouse and human (International Mouse Genome Sequencing Consortium, 2002). Comparative genomics is thus an attempt to take advantage of the information provided by the signatures of selection to understand the function and evolutionary processes that act on genomes.

The post-genomic era has contributed to our understanding in many scientific domains, including the study of gene expression regulation by Transcription Factors (TFs) through the identification of their cis-regulatory elements on promoter sequences. Here, comparative genomics plays an important role in the recognition of true functional Transcription Factor Binding Sites (TFBSs) that are predicted among a huge amount of false positives on genomic promoter sequences (see part 4.5). This method is based on the assumption that the functionally important regions, such as genes, exons or promoter elements, are evolutionarily conserved and can be discriminated from non-conserved ones that are supposed to be subject to fewer evolutionary constraints (Wasserman et al., 2000). Indeed, identifying and characterizing the TFBSs is a crucial step in the understanding of genomic regulatory regions and in the definition of the complex network of gene regulation. The elaboration of a comprehensive inventory of human regulatory motifs presents a major challenge aimed at providing a foundation for understanding cellular circuitry and its role in health and disease.

Chapter 2 - From gene to coding messenger in the post-genomic era

The path from gene to mature mRNA is a complex mechanism, highly regulated and giving the possibility of numerous variations at all levels: promoter, TSS (Transcriptional Start Site), alternative splicing etc. These variations multiply the codage ability of the gene and ultimately of the genome. So, the understanding of the regulation of eukaryotic gene expression is closely related to the precise understanding of the gene structures and products. A simplified model is widely used to summarize the path from the gene to the primary transcript and finally, to the coding messenger which will be subsequently translated into protein. Nevertheless, the in-depth study of a complex system such as the regulation of gene expression requires a deeper knowledge and a detailed description of each step.

Therefore, after a short presentation of the simplified model, we will present the crucial discoveries that provide more details of each step. This chapter will place particular emphasis on all the elements which play a role in gene expression regulation.

2.1 "Simplified" model

There are two general types of gene in the human genome: non-coding RNA genes and protein-coding genes. In the current manuscript, we will focus on protein-coding genes. As noted above, in eukaryotes, protein-coding genes exhibit incredible diversity in size and organization. There are, however, several conserved features that are generally admitted (Figure 4).

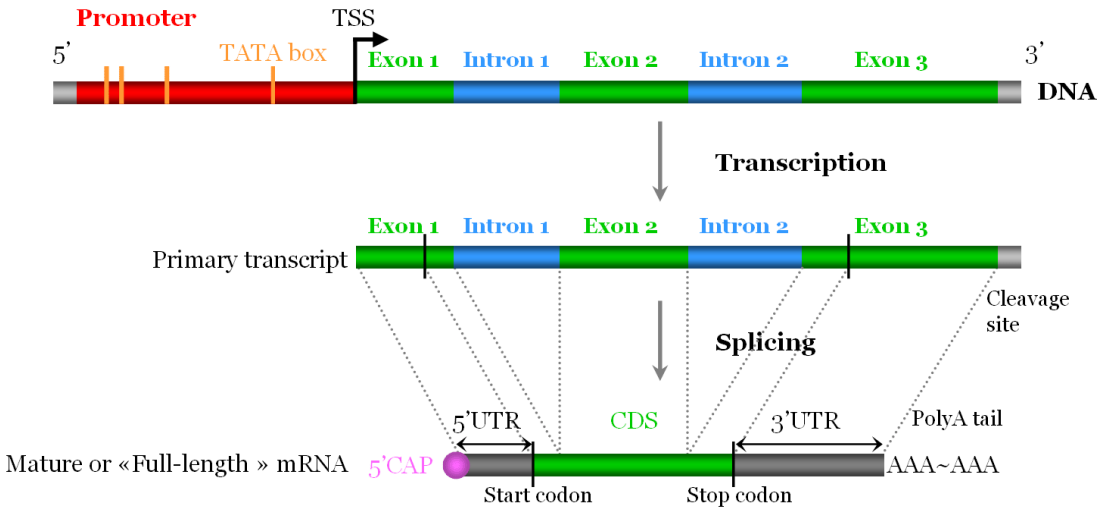


Figure 4. Classical eukaryotic gene to coding messenger model.

The promoter (in red) regulates the gene expression. The TSS determines the point of initiation of the transcription. Downstream the TSS, the gene is composed of alternated exons (green) and introns (blue). Indeed, most human genes are divided into exons and introns. The exons represent the segments that are present in the mature transcript (messenger RNA or mRNA), while the introns are removed from the primary transcript by a process called splicing. The core of the gene is the coding region or coding sequence (CDS, depicted in green in the mature or "full length" mRNA). It contains the nucleotide sequence that defines, and is translated into, the sequence of amino acids present in the protein. The coding region begins with the initiation or start codon (see Figure 4), which is normally ATG and ends with one of three termination or stop codons: TAA, TAG or TGA. On both sides of the coding region are noncoding sequences that are transcribed but are not translated. In the mature mRNA, these regions are called 5′ and 3′ UnTranslated Regions or 5′ and 3′ UTR (shown in dark grey). These non-coding sequences often contain regulatory elements that control mRNA stability and translation efficiency. Both the CDS and the UTRs may be interrupted by introns in the primary transcript (shown in blue).

The promoter (shown in red) can be defined as the gene region immediately upstream of the TSS and responsible for the regulation of the gene's expression. The function of the promoter is to integrate information about the status of the cell and to alter the rate of transcription initiation of a particular gene accordingly. The TATA box motif is mostly described in the promoter region.

Nevertheless, the post-genomic area has provided new insights which have led to a more complex model. This will be further detailed in the next section at four levels: the genome, gene, primary transcript and mature mRNA.

2.2 Model complexification

2.2.1 Genome level

In the context of gene regulation, it is important to first consider the genome level. Here, the different DNA packaging levels, the nucleosomal organization as well as the scaffold/matrix associated sequences or the repetitive sequences have been identified as playing a role in the regulation of gene expression.

2.2.1.1 DNA packaging

Genetic information is stored in the nucleus in the form of DNA. As the nucleus is limited in space and has to contain the billions of nucleotides that compose the genome, the DNA must be highly compacted and organized.

There are several levels of DNA packaging (Figure 5) from DNA double helix to chromosomes.

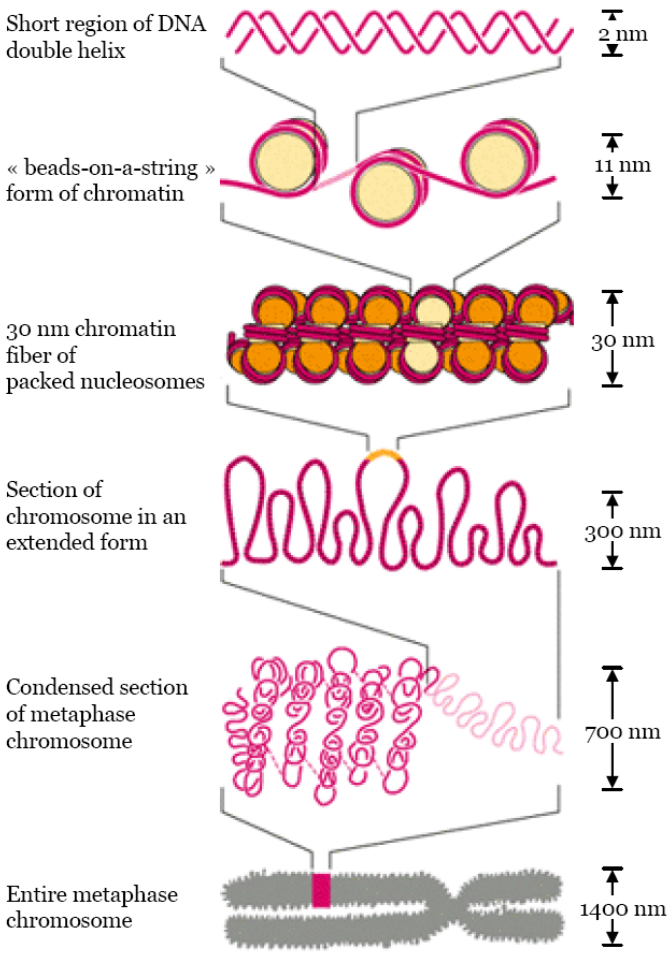


Figure 5. Different levels of DNA packaging.

The DNA starts out as single strand double helices and is condensed until chromosomal structures. From Cellupedia web server (http://library.thinkquest.org/C004535/chromosomes.html).

At the smallest level, the nucleotides are organized in the form of linear strands of double helices of 2 nm diameter. Zooming out, the DNA strand is wrapped around an octamer of histone molecules called a nucleosome (Figure 6).

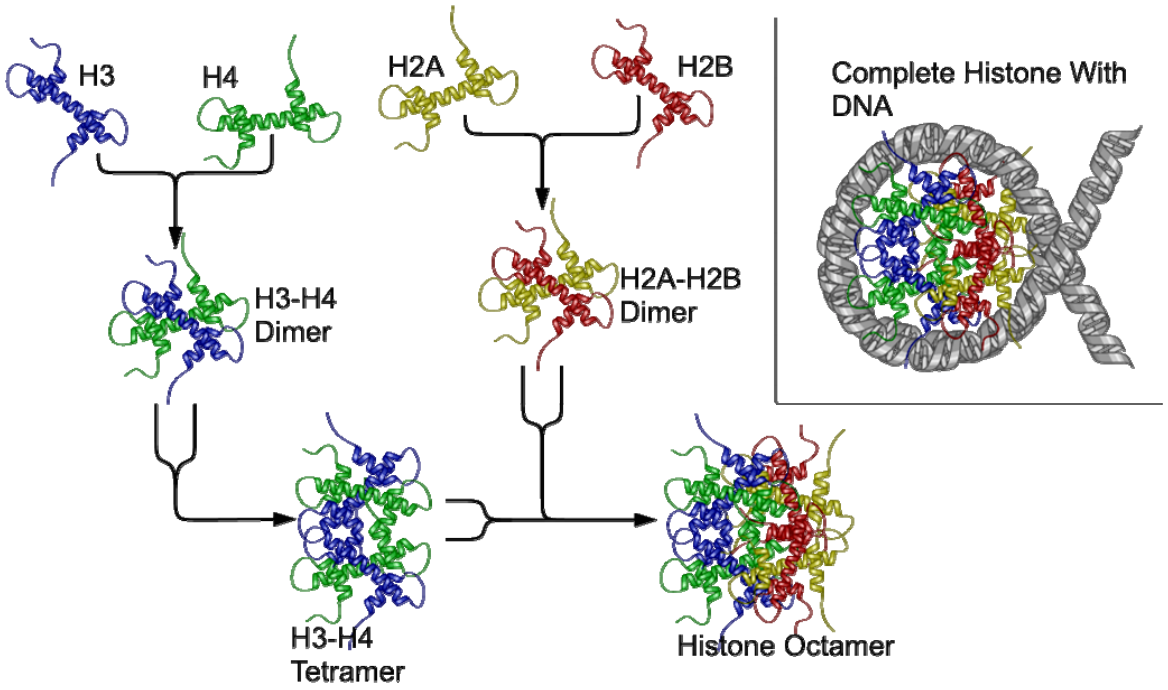


Figure 6. Schematic representation of the assembly of the core histones into the nucleosome structure.
This figure is referenced from "Molecular Biology of the Cell", Fourth Edition (Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter).

Two histones of each class H2A, H2B, H3 and H4 assemble to form a histone octamer, so-called core histones. Two turns of DNA, 146 base pair (bp) long, wrap around this protein core to form a nucleosome particle. The linker histone H1 (Figure 7) binds the nucleosome at the entry and exit sites of the DNA, thus locking the DNA into place and allowing the formation of higher order structure.

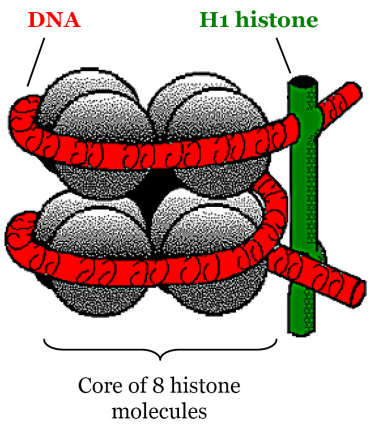


Figure 7. Binding of the linker histone H1 to the nucleosome (From http://www.accessexcellence.org/RC/VL/GG/nucleosome.html).

The most basic such formation is the 11 nm fiber or "beads-on-a-string" conformation. This involves the wrapping of DNA around nucleosomes with approximately 50 base pairs of DNA spaced between each nucleosome (also referred to as linker DNA). This nucleosome organization constitutes the first level of chromatin organization.

To further condense the DNA material, the linker histone H1 highly compacts the nucleosomes and stabilizes a higher order chromatin fiber of 30 nm diameter (Figure 5) that is fundamental to the structural organization of chromosomes. Indeed, the linker histones bind to each nucleosome, and by self affinity, link these nucleosomes together, resulting in the 30 nm chromatin fiber. The chromatin fibers then fold together into large looped domains which are organized into chromosomes during the mitotic cell cycle.

There are two types of chromatin: euchromatin and heterochromatin. The euchromatin is the lightly packed form of chromatin, the "beads-on-a-string" form of chromatin. This chromatin is often under active transcription. On the other hand, the heterochromatin is the tightly packed form of chromatin. This form is often characterized as "unactive" as the genes present in this chromatin are poorly expressed. Epigenetic alterations, such as DNA methylation and covalent histone modifications, notably histone acetylation, contribute to the regulation of gene expression by altering chromatin conformation.

Methylation of DNA is a common method of gene silencing. DNA is typically methylated by methyltransferase enzymes on cytosine nucleotides present in CpG islands which are rich in CpG dinucleotide sequence. Gene promoter CpG islands that acquire abnormal hypermethylation result in heritable transcriptional silencing. DNA methylation may impact the transcription of genes in two ways. First, the methylation of DNA may itself physically impede the binding of transcriptional proteins to the gene, thus blocking transcription. Second, and likely more importantly, methylated DNA may be bound by proteins known as

Methyl-CpG-Binding Domain proteins (MBDs). MBD proteins then recruit additional proteins to the locus, such as histone deacetylases and other chromatin remodelling proteins that can modify histones, thereby forming compact, inactive chromatin termed silent chromatin (Figure 8).

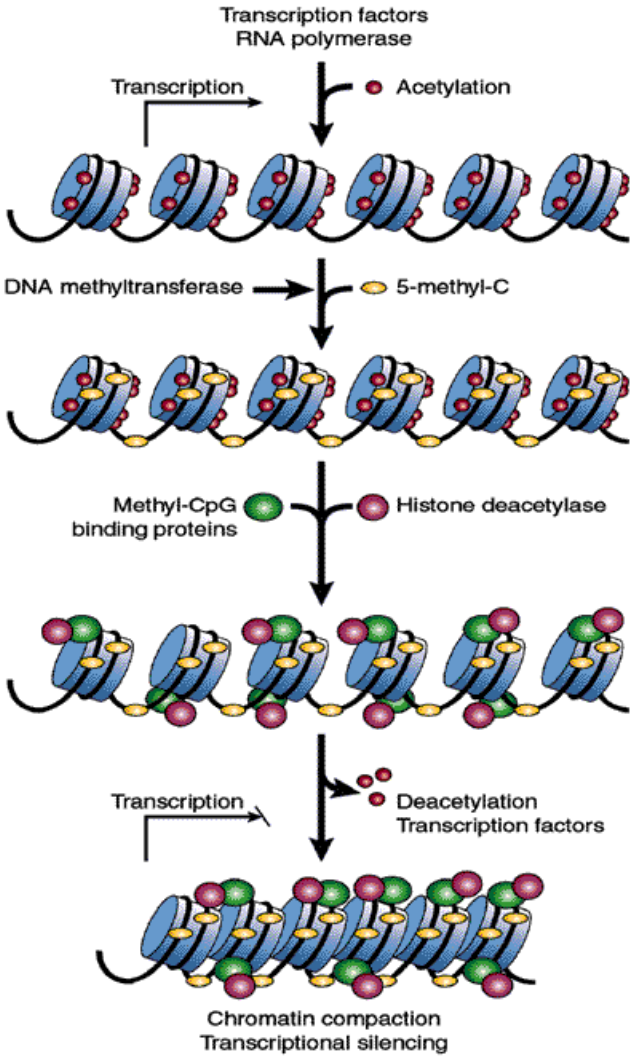


Figure 8. Transcriptionally active chromatin regions tend to be hyperacetylated and hypomethylated.
If a region of DNA or a gene is destined for silencing, chromatin remodeling enzymes such as histone deacetylases and ATP-dependent chromatin remodelers begin the gene silencing process. One or more of these activities may recruit DNA methyltransferase resulting in DNA methylation, followed finally by recruitment of the methyl-CpG binding proteins. The region of DNA will then be heritably maintained in an inactive state. Origin: http://www.med.ufl.edu/biochem/keithr/

DNA methylation and histone deacetylation often work together in gene silencing. The combination of these two mechanisms seems to be a signal for more compact DNA packaging, lowering gene expression (Figure 8).

If the octameric histone protein complexes largely determine the amount of supercoiling of DNA, their transcient modification (methylation, phosphorylation, acetylation…) clearly

plays a major role in transcription. Histone methylation is generally associated with transcriptional repression while histone acetylation is associated with transcriptional activation.

Histone AcetylTransferase enzymes (HATs) such as CREB (cAMP response element-binding protein)-binding protein (called CBP) supply a negative charge and neutralize the interaction of the N termini of histones with the DNA phosphate groups. As a consequence, the DNA is dissociated from the histone complex and thus the condensed chromatin is transformed into a transiently relaxed structure allowing transcription to proceed (Figure 9).

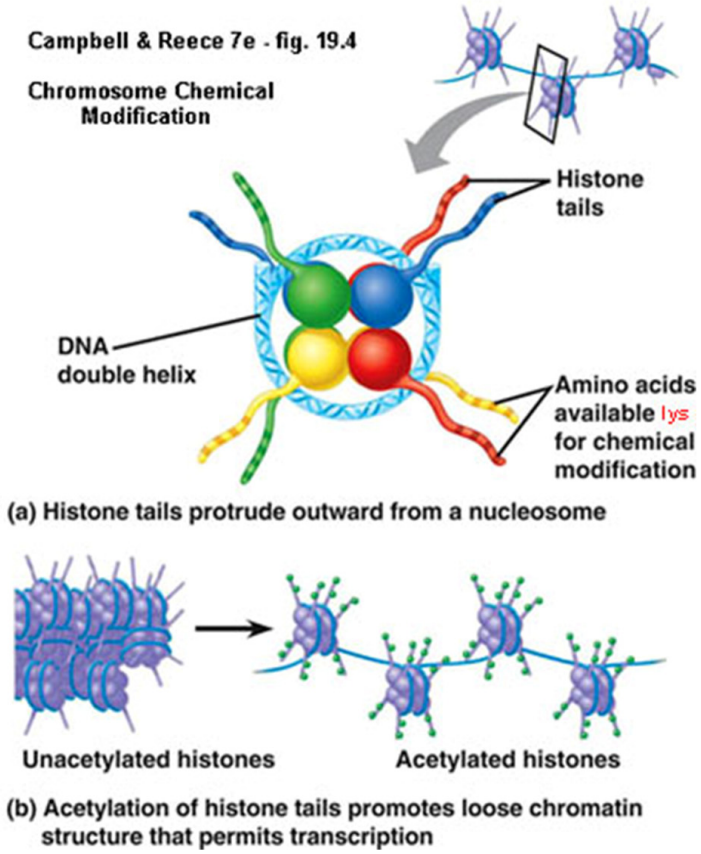


Figure 9. Histone acetylation is associated with active euchromatin.

DNA packaging through chromatin assembly and remodelling in the promoter regions of genes is implicated in eukaryotic transcription control especially for genes subjected to spatial and temporal regulation. The architecture of the chromatin in and around the gene plays a crucial role in the regulation of gene expression (Ganapathi et al., 2005). Furthermore, the chromatin environment of a region plays also a role in determining and modulating the expression status of its neighbouring genes (Grewal and Moazed, 2003).

2.2.1.2 Nucleosomal organization

As previously described, the primary template for local and global changes in the chromatin structure of a chromosome is the nucleosomal unit (Grewal and Moazed, 2003). Thus, nucleosomal organization over the promoter regions plays a major role in the regulation of the expression of downstream genes (Wolffe, 1994) (Khorasanizadeh, 2004).

Each nucleosome contains a 146 bp stretch of DNA, which is sharply bent and tightly wrapped around a histone protein octamer (Richmond and Davey, 2003). This sharp bending occurs at every DNA helical repeat (10 bp), when the major groove of the DNA faces inwards towards the histone octamer, and again 5 bp away, with opposite direction, when the major groove faces outward. Bends in each direction are facilitated by specific dinucleotides (Satchwell et al., 1986) (Widom, 2001). Neighbouring nucleosomes are separated from each other by 10–50-bp-long stretches of unwrapped linker DNA; thus, 75–90% of genomic DNA is wrapped in nucleosomes. Access to DNA wrapped in a nucleosome is occluded (Richmond and Davey, 2003) for polymerase, regulatory, repair and recombination complexes, yet nucleosomes also recruit other proteins through interactions with their histone tail domains (Jenuwein and Allis, 2001). Thus, the detailed locations of nucleosomes along the DNA may have important inhibitory or facilitatory roles (Kornberg and Lorch, 1999) (Wyrick et al., 1999) in regulating gene expression.

Recently, a eukaryotic nucleosome positioning code has been described (Segal et al., 2006). Indeed, nucleosomes have higher affinity for particular DNA sequences, reflecting the ability of the sequence to bend sharply, as required by the nucleosome structure. Segal *et al*, established a model that exhibits distinctive sequence motifs that recur periodically at the DNA helical repeat and are known to facilitate the sharp bending of DNA around the nucleosome (Widom, 2001) (Figure 10).

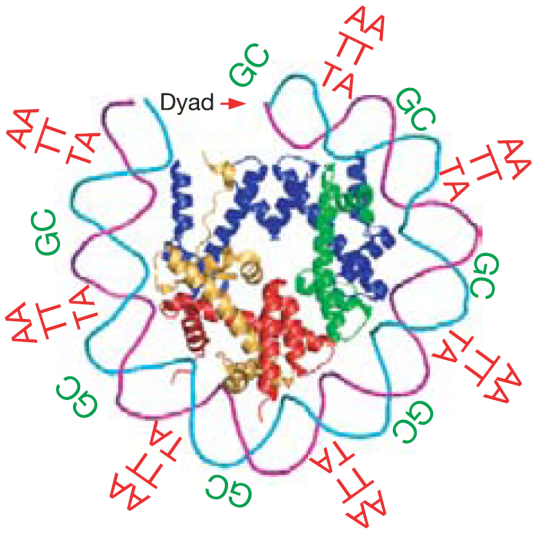


Figure 10. Nucleosome-DNA interaction model.
DNA dinucleotide sequence preferences of nucleosomes are represented relative to the three-dimensional structure of one-half of the symmetric nucleosome. Extracted from (Segal et al., 2006).

These sequence motifs include ~10 bp periodic AA/TT/TA dinucleotides that oscillate in phase with each other and out of phase with ~10 bp periodic GC dinucleotides. These distinctive motifs of the model represent DNA sequences that are preferentially occupied *in vivo* by nucleosomes and thus regulate the access for other proteins to DNA.

This intrinsic nucleosome organization may help in directing TFs towards subsets of appropriate target sites while excluding them from irrelevant sites. Of the 46 TFs tested in this study of Segal and coll. (2006), 37% had significantly lower nucleosome occupancy at their functional and conserved DNA binding sites (Segal et al., 2006) compared with the predicted occupancy at their canonical but presumed non-functional sites (Figure 11).

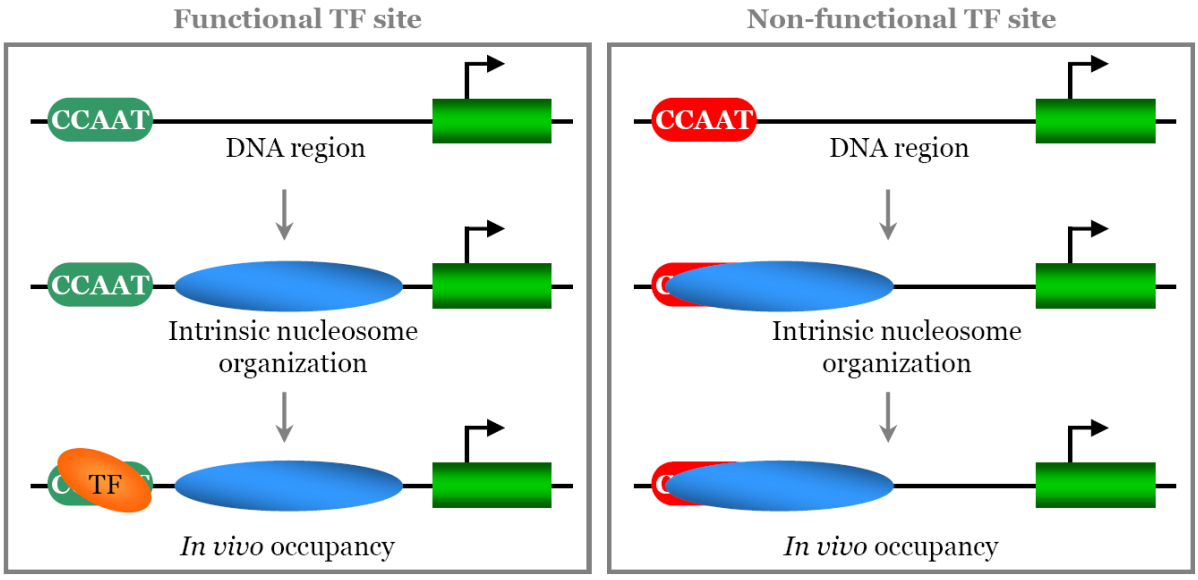


Figure 11. Nucleosome occupancy at TFBSs.
This intrinsic nucleosome organization may facilitate binding of transcription factors (TFs) at functional sites while disfavouring binding at identical non-functional sites that occur by chance. Extracted from (Segal et al., 2006).

The authors also found that the most probable location for TATA elements (Basehoar et al., 2004) correspond to genomic areas that remain unoccupied by nucleosomes; that is, just outside a stably positioned nucleosome. Therefore, eukaryotic genomes seem to direct the transcriptional machinery to functional sites by encoding unstable nucleosomes over these elements, thereby enhancing their accessibility.

Genomes encode an intrinsic nucleosome positioning code that can explain ~50% of the *in vivo* nucleosome positions. This nucleosome positioning code may facilitate specific chromosome functions including transcription factor binding, transcription initiation, and even remodelling of the nucleosomes themselves. This finding should help the elucidation of specific natural gene regulatory phenomena, such as the mechanism by which transcription factors (TFs) bind preferentially to appropriate sites in promoters rather than to the abundance of irrelevant sites in the genome.

So, the nucleosome distribution depends upon the occurrence of sequences destabilizing as well as forming nucleosome. Nucleosome destabilizing/excluding elements such as poly (dA.dT) and (CCGNN)n in promoter regions are implicated in maintaining constitutive gene expression (Suter et al., 2000). Nucleosome formation capability plays a role in the regulation of both housekeeping genes as well as tissue specific genes but with a reverse correlation (Ganapathi et al., 2005). Housekeeping genes show a significant enrichment in poly (dA.dT) stretches, which are known to destabilize nucleosomes. These genes indeed discourage nucleosome formation in order to match their expression profile in space and time by

ensuring enhanced accessibility to transcription machinery. Housekeeping genes generally have a coordinated expression and are clustered together in distinct chromatin domains of the chromosomes. In these chromatin domains, the presence of low nucleosome formation capability and enrichment of nucleosome destabilizing elements ensure an open chromatin configuration. On the other hand, tissue specific genes are known to be dispersed in heterochromatin regions having a high gene density (Versteeg et al., 2003) (de Laat and Grosveld, 2003). Tissue specific genes show high potential for nucleosome formation allowing selective accessibility to the transcriptional machinery. This is likely to be a local and restricted effect since the potentially distinct expression pattern of neighboring genes has to be preserved. It has been postulated that these genes protect themselves against the effects of positive and negative *cis*-acting elements of adjacent regions in order to maintain tissue specific expression profile. In this context, the boundary elements or the insulator model have been proposed to play a key role (de Laat and Grosveld, 2003) (see section 3.3).

2.2.1.3 Scaffold/matrix associated sequences

Scaffold/matrix attachment regions (S/MARs) are sequences that can attach themselves to the nuclear matrix, defined as a nuclear skeletal structure composed of a network of filaments linked to the nuclear lamina proteins. The S/MARs hence mediate the formation of independent dynamic chromatin loops (Glazko et al., 2001) that are known to be involved in transcriptional regulation of gene expression (Bode et al., 2000). This attachment of a DNA sequence to the matrix places the neighbouring genes in proximity of the TFs. The abundance of S/MARs in the 5' cis-regulatory regions of genes further demonstrates their role in transcriptional regulation (Glazko et al., 2003). S/MARs boundary elements are particularly enriched in the upstream regions of tissue-specific genes and might play a major role in facilitating tissue specific expression (Ganapathi et al., 2005). This matrix association potential of tissue specific genes would facilitate maintenance of functionally distinct domains to insulating themselves from both silencing and activating regulatory influence of adjacent domains.

2.2.1.4 Repetitive sequences

Repetitive sequences are implicated in chromatin organization and are predicted to play an important role in gene regulation (Grover et al., 2004) (Jordan et al., 2003) (Brahmachari et al., 1995). Interestingly, Alu elements have been shown to house TFBSs and the presence of such regulatory elements might influence chromatin structure and gene expression (Compe et al., 2005).

In general, the 5′ regions of both housekeeping and tissue specific genes have been shown to be enriched in short interspersed elements (SINES) in comparison to other classes of repetitive sequences. However, the differential distribution of repetitive sequences in these two classes of genes might be critical in maintaining distinct chromatin landscapes over these regions. The total repeat content in housekeeping gene regions is higher than in tissue specific gene regions and in particular, the 5' regions of housekeeping genes are more enriched in Alu sequences in comparison to those of tissue specific gene regions (Jordan et al., 2003) (Ganapathi et al., 2005).

Genes with high expression levels are clustered in genomic regions known as ridges. These gene rich regions are also characterized by a high (G+C) content, SINES and genes with short introns (Versteeg et al., 2003). Eisenberg and Levanon have reported the presence of significantly shorter introns and an overall compact gene structure in housekeeping genes as compared to non-housekeeping genes (Eisenberg and Levanon, 2003). The enrichment of SINES in the 5' regions of housekeeping genes suggests that they might be localized in the ridge regions of the genome. It has been suggested that the contrasting attributes of gene compactness, GC content and the length of the intronic and intergenic sequences might be involved in chromatin mediated regulation for maintaining distinct expression patterns in housekeeping and tissue specific genes (Vinogradov, 2004).

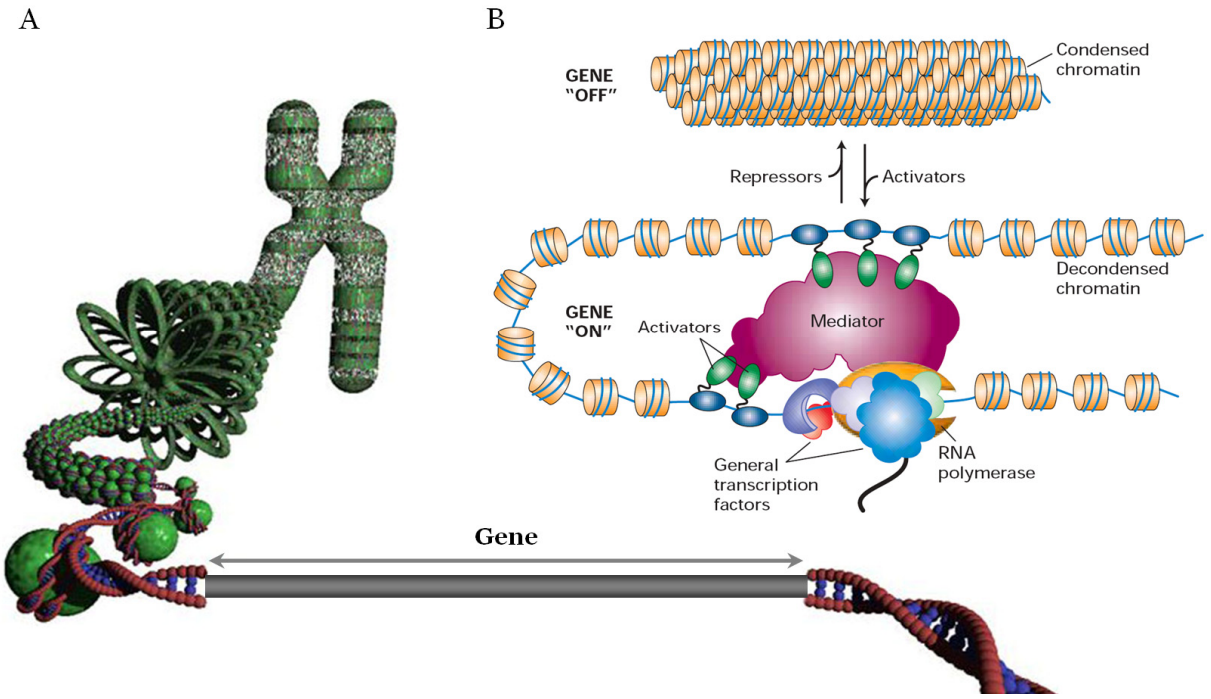


Figure 12. Model complexification at the genome level.
A- Gene integration in its genomic context. Extracted from http://www.gen.is/Erfdafraedi%20I/DNAsameindin.htm. B- The transcription control at the genomic level is mainly characterized by the chromatin structure with the condensed chromatin (heterochromatin) inactivating genes while the decondensed chromatin (euchromatin) is necessary for gene expression. Extracted from "Molecular cell biology", Lodish *et al.* fifth edition, Figure 11-1.

As a conclusion of the model complexification at the genome level (Figure 12), it is clear that nowadays, transcription regulation requires the integration of the gene in its genomic context in combination with "classical" transcription features, such as "preinitiation complex" or "loop structure", which are further described in the subsequent sections.

2.2.2 Gene level

2.2.2.1 Preinitiation Complex

In eukaryotic cells, transcription of every protein-coding gene is initiated by the assembly of an RNA polymerase II (pol II) preinitiation complex (PIC) on the promoter (Smale and Kadonaga, 2003). For many genes, it has been shown that the transcription level is regulated by controlling the efficiency of the formation of the PIC (Mitchell and Tjian, 1989) (Roeder, 1996). The PIC consists of the RNA polymerase II, the TATA-box Binding Protein (TBP) and a combination of six basal transcription factors (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TIIH)

which function collectively to specify the transcription start site (Reinberg et al., 1998) (Figure 13)

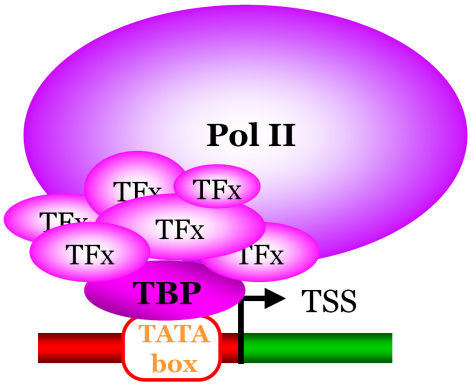


Figure 13. Polymerase II PreInitiation Complex recruitement.
The TATA-box Binding Protein (TBP), as well as the RNA polymerase II and other general TFs constitute the preinitiation complex (PIC) involved in the initiation of the transcription.

The PIC formation usually begins with TBP binding to the TATA box, initiator and/or Downstream Promoter Element (DPE) found in most core promoters (see section 3.2.1), followed by the entry of other general transcription factors and pol II. Formation and binding of this RNA polymerase II PIC to the core promoter surrounding the TSS of protein-coding genes is sufficient for a basal level of transcription (Thomas and Chiang, 2006).

2.2.2.2 Alternative Transcription Start Sites

Alternative TSSs are defined by alternative promoter sequences (Figure 14) and result in the transcription of alternate transcript isoforms.

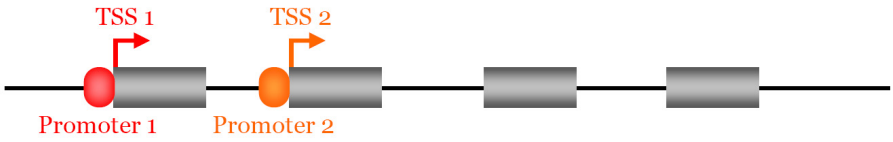


Figure 14. Schematic alternative promoter representation.

It has recently become clear that the phenomenon of alternative TSS is not as rare as previously thought but is rather a widely used mechanism. Preliminary analysis suggested that 18-20% of protein-coding genes use alternative promoters (Landry et al., 2003). Recently, the ENCODE (Encyclopedia of DNA Elements) consortium presented data based on a functional study of 1% of all human promoters (Cooper et al., 2006). This study revealed that more than 20% of genes have functional alternative promoters. However, the analysis of a larger data set revealed that 58% of protein-coding transcriptional units have two or more alternative promoters (Carninci et al., 2006). These recent analyses of Cap Analysis of Gene

Expression (CAGE) and other data have shown that most promoters have a wide distribution of closely located TSSs spanning a region wider than 50 bp. These results contrast with the "classical" view that defines most genes with one TSS governed by a TATA-box (Figure 4). These results lead to the conclusion that differentially regulated alternative TSSs are a common feature in protein-coding genes.

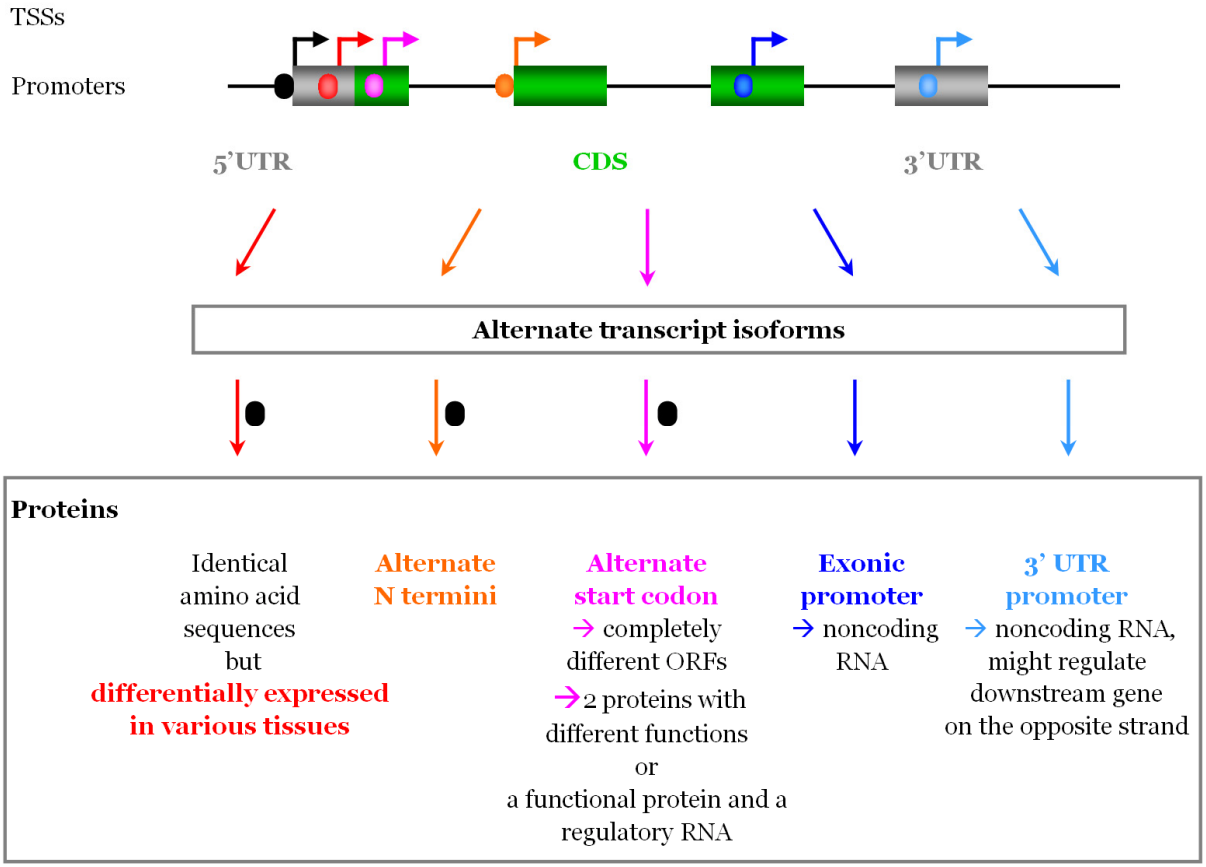


Figure 15. Schematic illustration of the different possibilities of alternative TSS on a gene and the consequences on the corresponding products. The black TSS is considered as reference for the cases depicted in red, orange and pink.

The different types of products resulting from alternative TSS are depicted in different colors in Figure 15 and are described in more detail below.

Functional alternative promoters can lead to alternative protein isoforms with or without significant similarity to each other. Alternative TSS can also generate proteins with strictly identical amino acid sequences (depicted in red in Figure 15). In these case, the alternative promoters function to provide distinct regulation of transcription for alternate RNA isoforms of the same gene. Indeed, most of the alternative isoforms have significantly different expression patterns.

Frequently, alternative promoters generate alternative N termini (case shown in orange in Figure 15) but some transcripts derived from alternative promoters can lead to an alternative

start codon in a different frame that results in a completely different protein (case depicted in pink in Figure 15) (Cooper et al., 2006).

In a recent study, high-throughput systematic 5′ end analysis of the mouse and human transcripts using the CAGE approach, allowed to define TSSs properties (Carninci et al., 2006). Typically, TSSs were identified over the known 5′end of the transcript and also in the 3′UTR. The TSS that occur between the 5′ end of the transcript and the 3′UTR, are mostly located in exons (depicted in dark blue and light blue in Figure 15). Such exonic promoter activity varies between genes and is conserved across species. The exonic transcription initiation sites might also have some relationship to exonic splicing enhancers and could have a role in RNA processing (Wu et al., 2005). In any case, the truncated transcripts generated from exonic promoters constitute a major class of noncoding RNAs (ncRNAs) whose actual functions are a matter of strong debates (Wang et al., 2004).

The promoter regions of these ncRNAs can contain specialized signals such as TFBSs and are generally more conserved than the promoters of the protein-coding mRNA, not only in mammals (e.g. between human and mouse) but also in vertebrates (down to chicken) (Cawley et al., 2004). A strong overrepresentation of three consecutive guanines, found at position -3 to -1 just before the 3′UTR TSS has recently been described (Carninci et al., 2006). Furthermore, an analysis of cross-species conservation between vertebrate genomes in the region surrounding the 3′UTR TSS revealed a high conservation of the region located at positions +40 to +90 relative to the TSS. Upstream regions of such 3′UTR TSS can initiate transcription. Transcripts initiated in 3′UTRs might function as regulatory noncoding RNAs of the downstream genes using a sense-antisense mechanism (see section 2.2.4.1), when downstream genes on the opposite strand are located much closer than expected (Carninci et al., 2005) (Katayama et al., 2005). Finally, it should be stressed that their transcriptional regulation is independent of the full-length transcript while the incidence of 3′UTR TSS is frequently tissue-specific.

2.2.2.3 General promoter structure

Briefly, the 5' segment immediately adjacent to the TSS includes the core promoter and the proximal promoter, which usually extends over about 200–300 nucleotides (Bortoluzzi et al., 2005). This region is involved in the modulation of transcription. The distal part of a promoter is variable in both composition and length, encompassing from 100 nucleotides to over 2 kb or even more. There is no clear-cut defined 5'-boundary for promoters (Werner, 1999). The detailed structure of higher eukaryotes promoter is presented in section 3.1 .

Promoters have a "block" structure which has been formed as a consequence of a selective pressure (Suzuki et al., 2004a). Suzuki *et al.*, studied potential promoter regions from – 1kb to + 200 bp with respect to the TSS designated as 0. These blocks have variable sizes with a mean length of 510 bp. Within these blocks, the sequence identity is uniformly 65% regardless of their length. Furthermore, most of the characterized TFBSs (Transcription Factors Binding Sites) are located within the blocks of conservation.

2.2.2.4 Transcription factors

Regulatory proteins, called TFs, modulate promoter activity and are one of the most important components in the transcriptional regulatory network. These *trans*-acting proteins control the rate of transcription at the level of the individual gene by binding crucial *cis*-regulatory elements or TFBSs on genomics sequences.

While the TFBSs are mainly described in the vicinity of the TSS, a variety of transcriptional regulatory elements are known to exist outside of the promoter regions of genes (Cooper et al., 2006). Indeed, they are present in introns, mostly the first intron, and in UTR regions.

Furthermore, although in simpler organisms, such as yeast, bacteria and viruses, TFBSs are usually associated with the promoters of their target genes, in more complex organisms, especially vertebrates, TFBSs are often positioned remotely from the genes they regulate—sometimes being as far away as a megabase from the TSS of a gene (Nobrega et al., 2003).

Control regions are modular in nature and expression of a given gene depends on specific combination of its regulatory elements and sometimes from their order and orientation (Bussemaker et al., 2001). Such modules of *cis*-regulatory elements are called *Cis*-Regulatory Modules (CRMs).

To illustrate how TFs can regulate the expression of genes, we will briefly describe a major class of the TFS: the nuclear receptors (NRs). NRs control a large number of physiological events. Signalling by nuclear receptors (NRs) is one of the major signal transduction paradigms invented by metazoans to regulate gene transcription (Robinson-Rechavi et al., 2003). Indeed, they constitute intracellular receptors which function as ligand activated transcription factors that can up or down regulate the expression of genes. NRs share a common structure, which consists of a transcription active domain, a DNA-Binding Domain (DBD) and a Ligand-Binding Domain (LBD) (Figure 16).

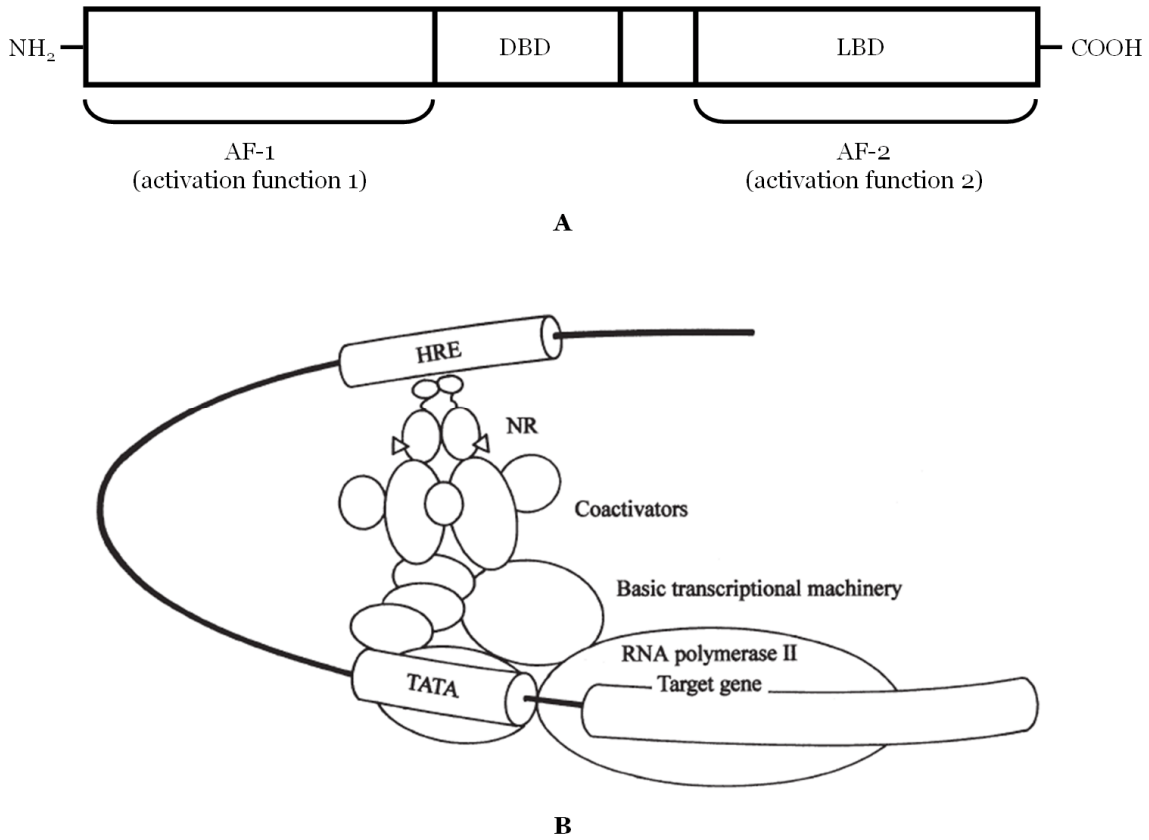


Figure 16. Nuclear Receptors
(A) Schematic structure of Nuclear Receptors (NRs). DBD, DNA-Binding Domain; LBD, Ligand-Binding Domain. (B) Schematic representation of the interactions among the NR, their DBD on the *cis*-regulatory module HRE (Hormone Response Element), some coactivators and the basic transcriptional machinery. Extracted from http://www.naika.or.jp/im2/43/05/05r.aspx.

Thus, their Ligand Binding Domain (LBD) can bind small hydrophobic molecules specifically. Molecular and structural studies reveal that these ligands constitute regulatory signals, which modify the NR transcriptional activity through conformational changes (Renaud and Moras, 2000). Furthermore, NRs recognize specific DNA binding sites by their DNA Binding Domain (DBD). NRs generally bind to DNA either as homodimers or as heterodimers. There are two transcription activation domains; the activation function-1 (AF-1) domain in the N-terminal region and the activation function-2 (AF-2) domain in the C-terminal region. While the AF-2 domain is relatively conserved among nuclear receptors, the AF-1 domain differs widely (Mangelsdorf et al., 1995) (Horwitz et al., 1996). When a ligand is bound to a receptor, the receptor changes in structure, translocates from the cytoplasm to the nucleus and then binds to the promoter region of the target gene. Coregulator proteins bind to the nuclear receptors and modulate the transcriptional activity of the nuclear receptors in a promoter- and cell-specific manner.

Taking into account all the previously described elements, we can define a more complex model at the gene level (Figure 17).

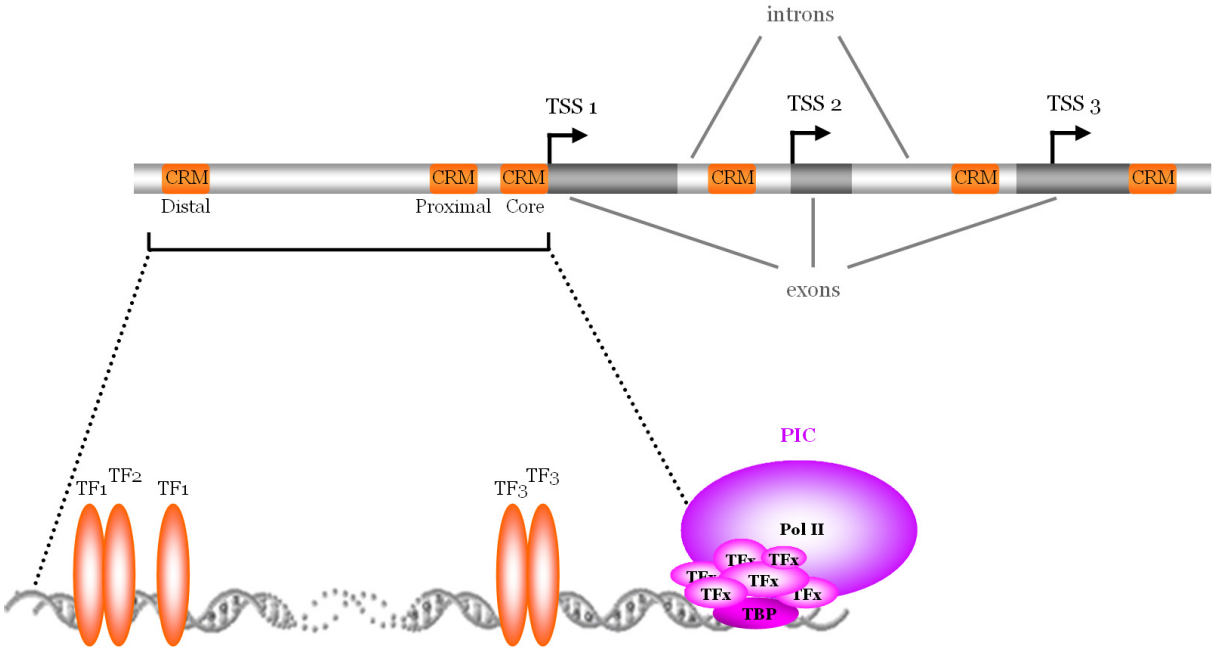


Figure 17. Complexification of the model at the gene level.
Extracted from http://www.uni-tuebingen.de/plantphys/harter/wanke/Projects_cis-Elements.html. Cis-regulatory modules (CRM) are described in the distal and proximal regions of the promoter, in introns and the UTR's. Alternative TSSs are described and imply different promoter regions.

2.2.3 Primary transcript level

The human genome sequencing project (Venter et al., 2001) has lead to the estimation of the total number of human genes as roughly 30,000, which is significantly less than the previous estimates of 100,000 to 150,000 genes based on analysis of expressed sequence tags (ESTs). This discrepancy can, in part, be accounted for by the alternative selection of promoters (alternative TSS, see section 2.2.2.2) and alternative mRNA splicing.

As previously pointed, mRNA splicing is an essential, precisely regulated post-transcriptional process occuring prior to mRNA translation. A gene is first transcribed into a pre-messenger RNA (pre-mRNA), which is a copy of the genomic DNA containing intronic regions destined to be removed during pre-mRNA processing (RNA splicing), as well as exonic sequences that are retained in the mature mRNA (see Figure 4). During RNA splicing, exons can either be retained in the mature message or targeted for removal in different combinations (pink lines in Figure 18) to create a diverse array of mRNAs from a single pre-mRNA, a process referred to as alternative RNA splicing (Lopez, 1998) (Figure 18).

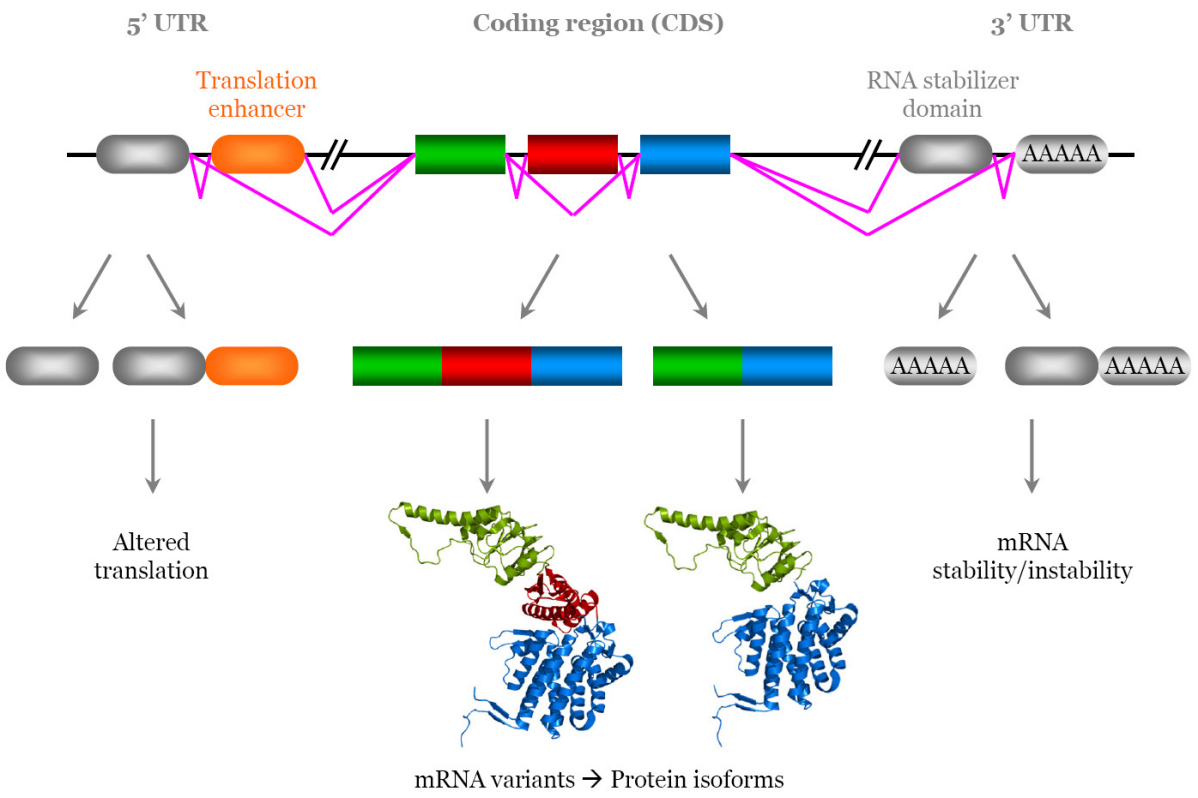


Figure 18. The impact of alternative mRNA splicing mechanism.

Alternative splicing events that affect the protein coding region of the mRNA will give rise to proteins isoforms which differ in their sequence, structure and therefore in their activities. Alternative splicing within the non-coding regions of the RNA (5' and 3'UTR) can result in changes in regulatory elements such as translation enhancers or RNA stability domains, which may have a crucial effect on the level of mRNA and on protein expression (see Figure 18).

2.2.4 Mature mRNA level

2.2.4.1 Antisense transcripts

The sense strand of DNA provides the template for production of mRNA, which in turn encodes proteins. Transcription from the opposite strand to a protein-coding or antisense (AS) strand can produce antisense transcripts. While 15-20% of protein-encoding genes have been identified with AS transcription (Okazaki et al., 2002) (Kiyosawa et al., 2003) (Yelin et al., 2003) (Werner and Berdal, 2005), large-scale cDNA sequencing in the FANTOM3 project (Waterston et al., 2002) suggests that antisense transcription is more widespread. For example, SAGE experiments indicate that at least 50% of all transcripts have a corresponding AS transcript in the brain (Siddiqui et al., 2005) while CAGE data show that up to 72% of the transcriptional units exhibit S/AS transcription (Gustincich et al., 2006).

Sense-antisense transcripts derive from overlapping genes on opposite strands sharing not only the same locus but also exonic fragments. The effects of eukaryotic natural antisense transcripts on the corresponding sense RNAs, their biological function and their involvement in physiological processes and gene regulation in living organisms are not fully understood. Nevertheless, number of documented examples indicates that they may exert control at various levels of gene expression, such as transcription, mRNA processing, splicing, stability, transport and translation. In many cases the complementary transcript does not encode for a protein product, but regulates expression by hybridizing to the mRNA of the controlled gene. Antisense transcripts can regulate gene expression by transcriptional interference, RNA masking and double-stranded RNA (dsRNA)-dependent mechanism (Lavorgna et al., 2004) (Munroe, 2004) (Noguchi et al., 1994).

In the case of the eIF2alpha protein, the transcription of the sense and antisense transcripts are inversely correlated (Noguchi et al., 1994). However, at other loci, sense and antisense transcription is not always observed to be mutually exclusive but the transcripts can be expressed in distinct tissue.

Formation of RNA duplexes between sense and antisense transcripts might mask some regulatory signals within either transcript and inhibit the binding of *trans*-acting factors. Indeed, the antisense RNA can specifically inhibit the alternative splicing of an mRNA, probably by blocking the accessibility of *cis*-regulatory elements.

There is evidence that interaction of antisense gene pairs can also affect gene expression through the activation of dsRNA-dependent pathways. This may include RNA editing or RNA interference (RNAi)-dependent gene silencing. Interaction between sense and antisense transcripts may result in the formation of dsRNA that might be modified by the RNA-editing machinery. Hyper-edited transcripts can be recognized by proteins, which inhibit their export from the nucleus and prevent translation. Alternatively, dsRNAs can be digested into small fragments (small interfering RNAs) by the RNA interference (RNAi) machinery (Kumar and Carmichael, 1998) (Sadiq et al., 1994). When dsRNAis are introduced into most eukaryotic cells, they are cleaved into 21-23 nucleotide duplexes. These fragments then target the specific destruction of cognate mRNAs (Hannon, 2002). The formation of double-stranded RNA to downregulate the expression of sense RNA molecule could make overlap quite hazardous. Any such double-stranded RNA is liable to be mistaken for viral DNA, leading to the destruction of the double-stranded RNA and homologous mRNAs by the cellular antiviral defense mechanism (Cullen, 2002). In extreme cases, formation of antiparallel heteroduplex RNA could completely block the expression of both genes. High-

level antisense transcription can inactivate a gene if the convergent transcript extends into that gene's coding region.

2.2.4.2 Post-transcriptional regulation by micro RNA

Post-transcriptional mechanisms affecting mRNA processing and stability play a role in regulating steady-state mRNA levels (Meyer et al., 2004) (Wilusz and Wilusz, 2004). Such mechanisms can involve regulation of the export of the mRNAs to the cytoplasm. However, we will focus on the effect of micro RNA as such a regulation can potentially be predicted *in silico* based on the DNA sequence.

MicroRNAs (miRNAs) have been widely studied and seem to have a relative importance in the regulation of the expression of the human genome (Xie et al., 2005). miRNAs are RNA genes that are the reverse complement of portions of another gene's mRNA transcript and inhibit the expression of the target gene. The 3'UTR regions are known to be involved in miRNA mediated regulation since they contain motifs involved in post-transcriptional regulation (Kuersten and Goodwin, 2003). Most of the 3'-UTR motifs have an 8-base length and a strong tendency to end with the adenyl nucleotide. These properties are reminiscent of a feature of miRNA (Bartel, 2004): many mature miRNAs start with a 'U' base followed by a 7-base sequence complementary to a site in the 3' UTR of the targeted mRNAs (Lewis et al., 2003). The binding of miRNA at such sites can guide degradation or repress translation of the mRNA. Nowadays, miRNA are estimated to regulate at least 20% of human genes. Most of these motifs seem to be conserved along the evolution. Nevertheless, in view of the high evolution rates of the miRNA, the 3'UTR 8-mers, strongly constrained in their evolution, may probably have additional roles than miRNA inhibition.

This chapter has clearly demonstrated the various aspects involved in the regulation of the expression of a target gene. Together, they contribute to the complexification of the "simplified" model aiming at a better understanding of this multifaceted biological process. The different aspects described here influence distinct levels in the regulation of gene expression, related to distinct biological fields of study. In this thesis, we focus more particularly on the promoter and *cis*-regulatory module aspects.

Chapter 3 - Promoters of mammalians

Tens of thousands of mammalian genes are expressed in various cells at different times and development stages, controlled mainly at the promoter level through the interaction of Transcription Factors (TFs) with *cis*-regulatory elements. The Transcription Factor Bindings Sites (TFBSs) and promoters were introduced briefly in Chapter 2 -. As they are central to this work, this chapter is dedicated to a more comprehensive description of TFBSs and promoters. First, their composition and structure are detailed. Next, several characteristics of promoters are presented. Finally, the conservation of TFBSs and promoters during evolution are discussed.

3.1 Transcription Factor Binding Sites (TFBSs)

TFs play a central role in the transcriptional regulation of gene expression by binding to specific, short DNA sequence motifs, called TFBSs. The TFBSs are located in regions involved in the regulation of gene expression, the promoter regions (see 2.2.2.3). In this section, the properties of TFBSs and their organization in *Cis*-Regulatory Modules (CRMs) in these promoter regions are presented.

TFBSs are short DNA sequences, described as being from 6 to 10 (Wray et al., 2003) or from 5 to 25 nucleotides long (Qiu, 2003). TFBSs recognized by the same TF can be found in different promoters. Furthermore, TFs can tolerate widely variable target sequences (Thompson et al., 2004b).

The binding of a TF is not simply a function of the theoretical affinity for a DNA site, but also of a number of other factors such as the chromatin environment and the cooperation or competition with other DNA-binding proteins. In higher eukaryotes, TFs rarely operate by themselves, but rather bind to DNA in cooperation with other DNA-binding proteins (Blanchette et al., 2006). In fact, there is considerable evidence that TFBSs which direct gene expression occur in clusters of 50-200 nucleotides in higher eukaryotes constituting CRMs (Arnone and Davidson, 1997) (Figure 19), in which the weak individual signals provide a collectively strong signal. These CRMs have been the focus of many studies recently, particularly in the context of the gene regulation during development (Howard and Davidson, 2004), and are believed to be key features of most transcriptional regulatory

processes in mammals. The CRM can be viewed as a circuit translating input signals from diverse pathways into an output, gene activity, through the combinatorial interaction of multiple TFs with target *cis*-regulatory units.

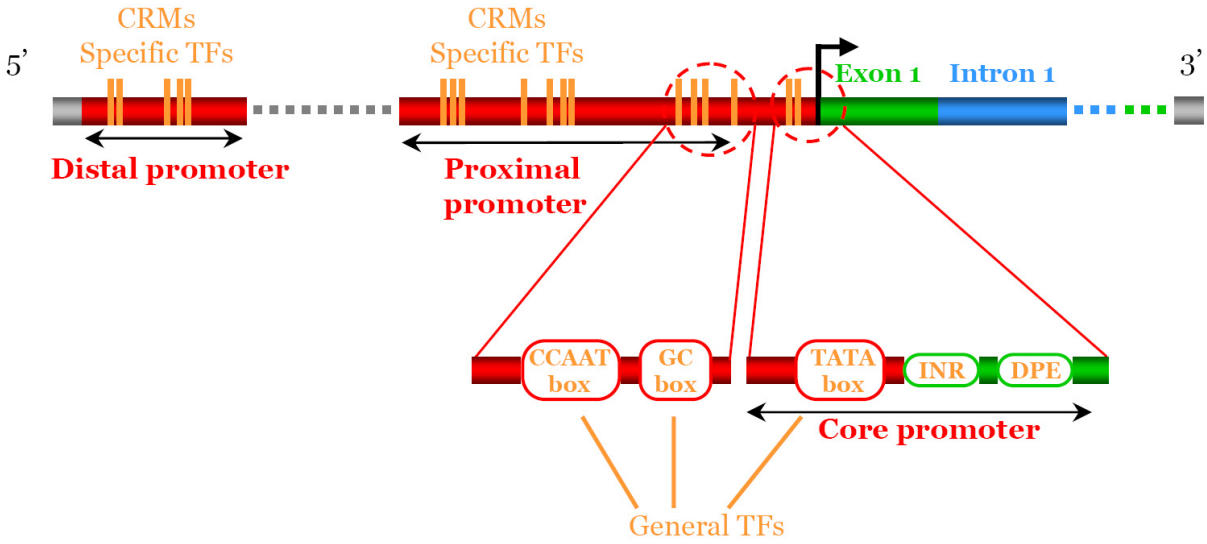


Figure 19. Detailed promoter structure of higher eukatyotes.
The TFBSs clustered in *cis*-regulatory modules (CRMs) are depicted in orange in the core, the proximal and the distal promoters. General TFs recognize TFBSs in the core promoter and in the proximal promoter region close to the core promoter, while specific TFs bind clusters of TFBs in the proximal and the distal promoter.

Several features characterize the known CRMs: (1) CRMs are generally composed of several binding sites for a few different TFs; (2) CRMs, and in particular the binding sites they contain, are generally more evolutionarily conserved than their flanking intergenic regions, and (3) genes regulated by a common set of TFs tend to be coexpressed (Blanchette et al., 2006).

TFBSs are often overrepresented and appear in multiple copies inside the CRMs to form cooperating units (Cora et al., 2005). Furthermore, a defined order of multiple TFBSs in the CRMs has been shown to be critical for modulating the expression of genes (Christoffels et al., 1998) (Klingenhoff et al., 1999). For example, it has been shown that for proper spatial expression in the endoderm of the sea urchin, one particular pairing of Gata sites is essential and that these function synergistically with an adjacent Otx site (Yuh et al., 2004).

The number of distinct TFs binding a module, the number of TFBS, the spatial constraints, the order of TFBS and relative strands of TFBS differ for different regulatory pathways (Ovcharenko and Nobrega, 2005). It has been previously shown that in several eukaryotes including yeast, worm and flies regulatory elements with similar function operate under

similar organizational principles, the modular distribution of a defined set of TFBSs. Recent evidence suggests that this could also be the case in humans (Donaldson et al., 2005).

As described above, TFBSs, organized in CRMs, play a major role in the regulation of gene expression and are frequently present in the promoter of the gene. The promoter itself (Figure 19) is structured in distinct regions (core promoter, proximal promoter and distal promoter) that will be described in more detail in the subsequent sections.

3.2 Core and proximal promoters

The core and proximal promoters are often described as a single region, but here we will present them individually in order to distinguish between the two elements.

3.2.1 Core promoter

In eukaryotes, the core promoter serves as a platform for the assembly of the transcription PIC. In fact, this region constitutes the ultimate target of the vast network of regulatory factors that contribute to the initiation of transcription by RNA polymerase II (see Figure 19). Generally, this recognition region for the basal transcription apparatus is located in the vicinity of the TSS. More precisely, in the human genome, the motifs involved in transcription initiation preferentially occur within ~ 50 to 100 bases straddling the TSS (Xie et al., 2005) (Lee and Young, 2000). Nevertheless, the core promoter has also been defined as compact and composed of up to ~35 bp from the TSS or ~60 bp straddling the TSS (Levine and Tjian, 2003).

Several *cis*-regulatory sequence elements are embedded in the core promoter, they help to direct and orient the PIC to this region (2.2.2.1). These *cis*-acting elements are recognized by the RNA polymerase II and by general and basal TFs. Their recruitement to the promoter accelerates or inhibits the formation of the PIC through direct interaction or by changing the conformation of the DNA (Novina and Roy, 1996).

There are at least three different sequence elements that can recruit the TBP containing TFIID initiation complex: the TATA-box, the initiator element (INR), which is a Pol II binding site, and the Downstream Promoter Element (DPE) (Smale and Kadonaga, 2003) (Figure 20).

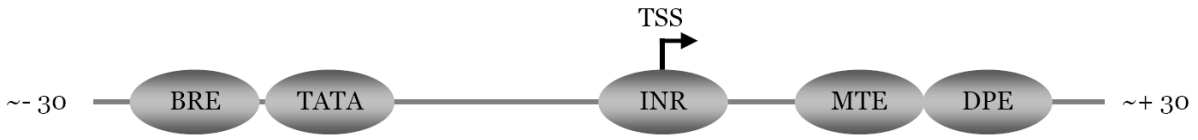


Figure 20. Schematic diagram of the core promoter elements.

They constitute directional motifs and strand specific core promoter elements which are present on the positive strand. Other motifs are non-directional motifs and occur on both strands (Fitzgerald et al., 2006).

The main function of the TATA box is to anchor the PIC to guide RNA polymerase (Kadonaga, 2002) ahead of the TSSs (Figure 13).

The INR spans the TSS. It can function independently or in combination with a TATA box, exerting a positive effect on transcription. The sequence of INR has recently been characterized (Carninci et al., 2006). The transcription preferentially starts with a pyrimidine at position -1 followed by a purine at position +1. The most preferred initiator dinucleotides are CG, CA and TG; they are associated with more active TSSs. The INR sequence is commonly subject to evolutionary changes among mammals.

The DPE occurs mostly in TATA-less promoters (Jin et al., 2006) (Burke and Kadonaga, 1996). The TATA-box and the DPE have symetrical positions relative to the TSS and similar function implying their mutual exclusion. Thus, DPE is structurally and functionally analogous to a TATA-box (Burke et al., 1998).

Additional motifs such as the Motif Ten Element (MTE) and TFIIB Recognition Element (BRE) have also been described as being associated with core promoters and are important for transcription by RNA polymerase II (Lim et al., 2004) (Deng and Roberts, 2005). The MTE requires and cooperates with the INR to stimulate transcription (Lim et al., 2004) but functions independently of the TATA-box and DPE. BRE orients the PIC on the promoter.

However, these elements are not present in all promoters (Smale and Kadonaga, 2003) (Hahn, 2004). In fact, no known DNA sequence motifs are shared by all core promoters.

Cooper *et al.* have developed a comprehensive functional testing of DNA fragments likely to be promoters covering 1% of the human genome (Cooper et al., 2006). They reported that, on average, the sequence -300 to -50 bp of the TSS contributes positively to core promoter activity. In 68% of the cases, the presence of 40 bp upstream of the predicted TSS is sufficient to maintain basal activity. Furthermore, these regions contain much of the constraint observed in promoters. These observations clearly emphasize the importance of the core promoter.

As more and more promoters are functionally characterized, the concepts of the "general transcription machinery" and "basal promoter elements" will be continuously refined.

3.2.2 Proximal promoter

The proximal promoter is located upstream of the core promoter up to roughly -250 bp from the TSS (Figure 19). It contains *cis*-regulatory sequences that are recognized by TFs specific to cell or tissue types and that control spatial and temporal expression of the downstream gene. Indeed, the proximal promoter, as well as the enhancers, silencers and insulator elements (see section 3.3), contain recognition sites for a variety of TFBSs (Butler and Kadonaga, 2002).

Most of the TFBSs in the promoter show tissue specificity and distance constraints with respect to the TSS (Carninci et al., 2006). In addition, some motifs tend to appear in multiple copies defining CRMs in the studied proximal promoter (Xie et al., 2005) with regulatory elements controlling transcription rates.

Proximal promoters may not contain all the information required to precisely control transcription of individual genes in time and space. The regulatory factors of the proximal promoter do not always function as classical activators or repressors; instead, they might serve as 'tethering elements' that recruit distal enhancers to the core promoter (Su et al., 1991) (Calhoun et al., 2002).

3.2.3 Statistics on specific motifs

Most core promoters have long been thought to contain specific motifs such as CpG island, INR, TATA-box, GC-box and CAAT-box and DPE. However, more and more evidence has been provided that most human core promoters do not have all these elements and notably the well studied TATA-motif (Florquin et al., 2005) (Fukue et al., 2004). In fact, it is becoming clear that only ~ 10% of promoters may contain TATA-box (Cooper et al., 2006). In accordance with these results, the TATA box has not been found to be significantly enriched in the promoters studied by Kim *et al.* (Kim et al., 2005) (Figure 21).

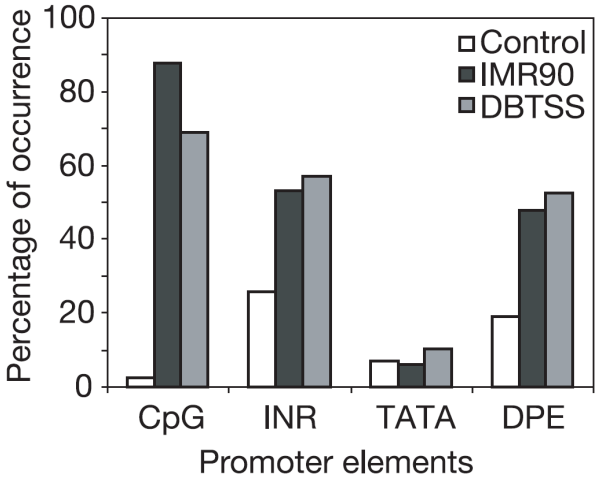


Figure 21. Presence of CpG island, INR, TATA-box and DPE in human promoters.
Chart showing the percentages of active promoters detected in IMR90 fibroblast cells or DBTSS active promoters overlapping with CpG islands, or containing conserved TATA box, INR or DPE elements. (Extracted from Kim *et al.* (Kim et al., 2005)).

The previous finding of over-representation of core-promoter motifs might have been biased because in general these studies were performed on very small and specific datasets (Lagrange et al., 1998) (Kadonaga, 2002) (Bucher, 1990). Nevetheless, the results obtained from high-throughput genome-level analyses, although sometimes controversial, are defining a new landscape for the core and proximal promoters.

Many TFBSs such as the TATA, GC and CAAT boxes often occur near the TSS (Suzuki et al., 2001). Most known TFBSs in TRANSFAC (see 4.3.2) have preferred locations between -300 and +50 bp relative to the TSS (Marino-Ramirez et al., 2004). This finding suggests that the basal promoter and nearby upstream regulatory elements are found in the region between -300 and +50 bp, in agreement with a recent study from the Myers laboratory, where 91% of 152 DNA fragments containing regions -550 to +50 relative to the TSS were active as promoters in at least one of four cell types evaluated (Trinklein et al., 2003).

TATA-box promoters are associated with sharper distinct TSSs with a single dominant TSS, tightly regulated transcripts, tissue-specific genes and high conservation across species. CCAAT-box and GC-box sequences that are not associated with CpG islands are preferentially associated with such single dominant TSS. More than 70% of TATA-boxes are located within 50 bp upstream the TSS between position -33 and -28, with preferred positions -31 and -30 (Carninci et al., 2006). This spacing can be explained by structural constraints showing that the distance from the TATA-box to the active center of the polymerase II is 30 bp (Smale and Kadonaga, 2003). In the analysis of transcriptional promoter structure and function in 1% of the human genome (Cooper et al., 2006), CAAT (CCAAT) boxes (19% of the functional promoters) were even more frequent than TATA-boxes (TATA(T/A)(T/A)) (16% of

the functional promoters). However, no significant correlation has been found between the presence of CAAT and TATA boxes and promoter activity (Cooper et al., 2006) (Trinklein et al., 2003). The 3D structure of the core promoter may be even more important for the initiation of transcription than the presence of the TATA-motif, or other motifs such as a binding site (Leblanc et al., 2000).

Several definitions of CpG islands have been proposed. The strict NCBI definition considers a DNA region as being a CpG island, if (i) it is >500 bp in length; (ii) it has a G+C content >50%; and (iii) its ratio of observed/expected CpG is greater than or equal to 0.6 (http://www.ncbi.nih.gov/mapview/static/humansearch.html#cpg). Marino-Ramirez *et al.* (Marino-Ramirez et al., 2004) have studied a set of 4737 distinct Putative Promoter Regions (PPRs) from the human genome. While previous studies estimated that 40±50% of human genes overlap with CpG islands (Suzuki et al., 2001) (Larsen et al., 1992), they found that 76% of their PPRs (3,608 of 4,737) overlap with at least one CpG island. These results are in agreement with those of Kim *et al* (Kim et al., 2005) who also showed that CpG-associated promoters (88%) (Figure 21) are significantly higher than previous estimates (56%, (Antequera and Bird, 1993)). This suggests that CpG islands might play a more general role in gene expression than previously appreciated.

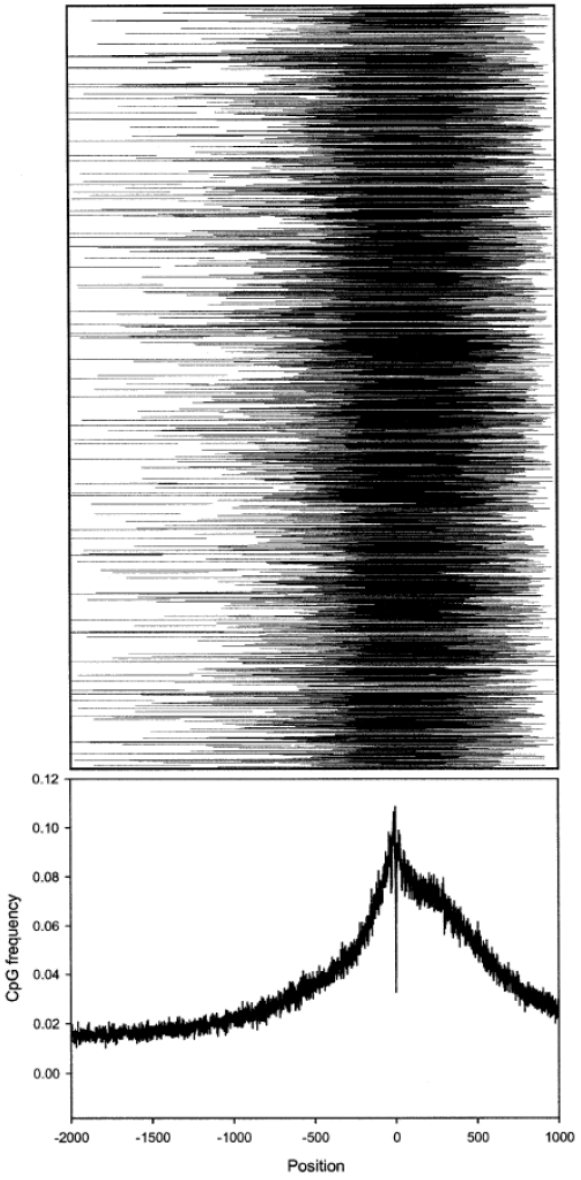


Figure 22. CpG frequency in the promoter regions.
CpG Islands in the putative promoter regions (PPRs) are depicted in the upper panel. The CpG islands are represented as black lines in the promoter regions while other regions are represented as white spaces. Positional CpG frequency in the promoter regions is illustrated in the lower panel. The nucleotide positions are relative to the TSS (position 0). Extracted from the study of Marino-Ramirez (Marino-Ramirez et al., 2004).

As shown in Figure 22, CpG islands appear to cluster near the TSS, but in many cases they extend over longer regions (Marino-Ramirez et al., 2004). Furthermore and consistent with previous results, Marino-Ramirez *et al.* highlighted that 28.4% of the promoter regions contained TATA boxes; 88.6% GC boxes; and 60% CAAT boxes. They also revealed possible regions preferred for transcription factor binding. Dinucleotides containing C or G are overrepresented in the promoter, whereas dinucleotides containing T or A are under-represented. The CpG frequency generally increased with proximity to the TSS, but decreases dramatically at positions ±29 to ±24 and ±1. Probably, the presence of TATA boxes

in this region causes the decrease, because they increase the frequency of the dinucleotides TpA, ApT and ApA at these positions. This characterization of dinucleotide compositions near the TSS is in general agreement with previous observations (Down and Hubbard, 2002).

As CpG islands often reside in the promoter region of genes, and the methylation of CpGs in these regions is thought to affect the expression of their downstream genes. For example, several transcription factors are known to exhibit differential activities depending on whether or not their *cis*-elements are methylated (Cross and Bird, 1995) (Costello et al., 2000). In addition, CpG methylation may have a role in the conservation of CpG islands whigh are thought to have been depleted during evolution by the transition mutation from methylated cytosine to thymine (Bird and Taggart, 1980). However, CpG islands are still found in several vertebrate genomes (Gardiner-Garden and Frommer, 1987), and it has been reported that the dinucleotides in CpG islands are usually unmethylated, thus avoiding the above mutations (Larsen et al., 1992).

Gardiner-Garden and Frommer (Gardiner-Garden and Frommer, 1987), and Larsen *et al.* (Larsen et al., 1992) first examinated the effect of CpG islands on the tissue specifity of nearby gene expression. They reported that promoters of most housekeeping genes contain CpG islands while promoters of tissue-specific genes lack them. However, their analyses were based on the very small subset of genes available in those days. GenBank-based sequence analyses have shown that heavy (GC-rich) isochors tend to contain housekeeping genes, while light (GC-poor) isochors are rich in tissue-specific genes (Pesole et al., 1999). A genome-wide analysis based on the SAGE data also indicates that housekeeping genes tend to be found in regions of high GC content (Versteeg et al., 2003). Since the CpG islands tend to exist in heavy isochors, this seems to be in good agreement with the previous results. Yamashita *et al.* have shown that an evolutionary conserved tendency exists that a gene lacking CpG islands around its TSS is expressed with a higher degree of tissue specificity (Yamashita et al., 2005).

However, they are also some reports in which no correlation was found between GC content and tissue specificity. For example, no significant expression difference was observed between light and heavy isochores in studies based on the EST data (Goncalves et al., 2000) (Ponger et al., 2001). Moreover the correlation may not be evolutionary conserved. Indeed, Vinogradov reported that there is a correlation between the GC content of the third position of codons and their tissue specificity in human, but not in mouse (Vinogradov, 2003).

CpG rich promoters have been shown to generally have broad TSS regions with no single dominant TSS and to be associated with ubiquitous housekeeping trancripts. One exception is the central nervous system-specific promoters, which are especially CpG-rich (Carninci et

al., 2006). The regulation and the promoter region of brain specific genes seem to be unusual. Indeed, genes specifically expressed in brain are surprisingly enriched in CpG islands, compared to other tissue-specific transcripts (Gustincich et al., 2006). Due to the nature of the broad CpG promoters, it is likely that brain-specific transcription is regulated at the epigenetic level much more frequently than tissue–specific expression in other organs. Nevertheless, although they contain CpG islands that are fast evolving, brain-specific promoters have a high conservation over the -300 to +50 promoter region (Gustincich et al., 2006). This highlights another exception: brain-specific promoters are not under evolutionary selection. It is unclear why the brain has different RNA expression regulatory mechanisms. Nevertheless, efficient transcription may be linked to the transcription factor Sp1 that can recruit the TATA-binding protein in the absence of TATA-boxes (Butler and Kadonaga, 2002). Indeed, Sp1 sites are overrepresented in promoters associated with CpG islands (Carninci et al., 2006).

3.2.4 Core and proximal promoters classification

Several ways to classify mammalian promoters have been proposed either using functional or compositional properties.

Classically mammalian promoters are separated into two classes according to their sequence context (Carninci et al., 2006). The first class as been described as containing the conserved TATA box promoters. They have the ability to initiate at a well defined site. The second class corresponds to the more plastic, broad and evolvable CpG-rich promoters. As previously discussed, the TATA-box promoters represent a minority of mammalian promoters, the majority being CpG islands promoters.

Kadonaga *et al.* (Kadonaga, 2004) used the presence of functional core promoter elements such as TATA boxes, initiators, and downstream promoter elements to classify promoters. A different approach was used by Kim *et al.* (Kim et al., 2005): the properties of preinitiation complex binding to the promoter and the observed transcript expression state allowed them to reveal four general classes of promoters that define the transcriptome of the cell.

Alternatively, Yamashita *et al.* (Yamashita et al., 2005) classified human and mouse genes based on the presence of a CpG island within the -100 to +100 region while Bajic *et al.*, (Bajic et al., 2006) also studied the nucleotide content within the [-100, +100] bp relative to the TSS and defined four distinct types of TSSs, from A to D. The Table 4 describes the GC content of these four TSS types.

TSS Type	Upstream GC Content	Downstream GC Content
A	GC-rich	GC-rich
B	GC-rich	AT-rich
C	AT-rich	GC-rich
D	AT-rich	AT-rich

Table 4. Four TSS types based on the GC content upstream and downstream of the TSS. GC-rich means more than 50% of GpC in the considered region. AT-rich (i.e., GC-poor) means less than 50% pf GpC in the considered region. In this case, the upstream region is [-100, -1], and the downstream region is [+1, +100] relative to the TSS.

Many of the TSSs that are not evidently GC-rich (TSS types B to D) have changing GC content when going from upstream to downstream regions (Figure 23).

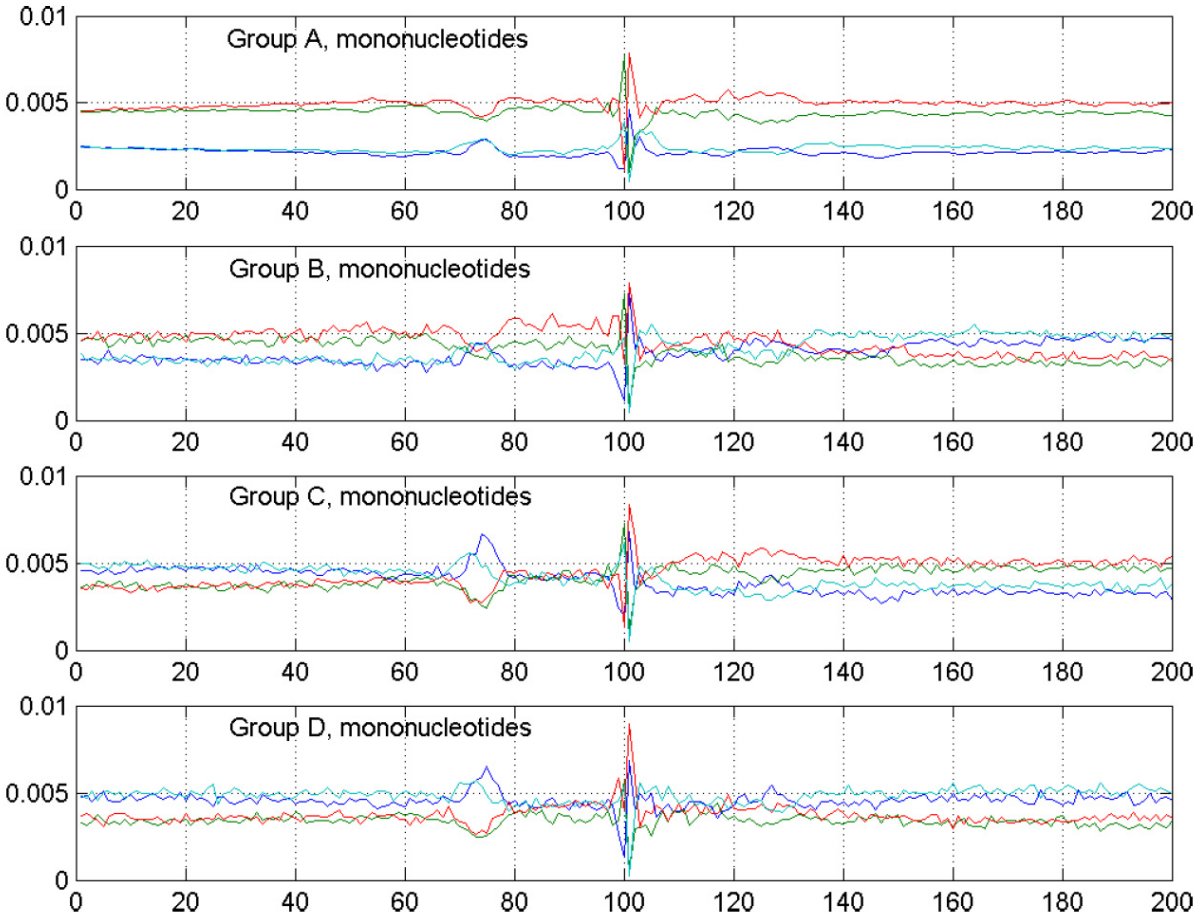


Figure 23. Distribution of mononucleotides in mouse promoters in the region surrounding the TSS (position 100).
The nucleotides adenine, cytosine, guanine, and thymine are represented by blue, green, red, and light blue, respectively. The TSS types that are GCpoor upstream (C and D) show very characteristic enrichment in adenine and thymine nucleotides around [-35,-20], suggesting a potential dominant influence of TATA box and similar AT-rich elements in transcription initiation in these types. In type B and A TSSs, this influence does not seem to be dominant, but the presence of such elements is suggested by a significant reduction of the GC content in the [-35,-20] region. In principle, one could attempt to link the types of AT-rich upstream elements with initiating dinucleotides characteristic of different TSS types. Extracted from (Bajic et al., 2006).

Furthermore, some eukaryotic genomes have dominant TSS characteristic of one of these classes. For example, TSS types B and C are prevalent in *Takifugu rubripes* and type D in *Drosophila melanogaster*, while type A is characteristic of the human genome (Aerts et al., 2004). For all TSS types the number of enriched TFBSs in the upstream region is much higher than in the downstream region. In three TSS types (A, C and D) the number of TFBSs in the downstream region is minimal compared to the upstream region. The only exception is type B, for which there are a significant number of enriched TFBSs in the downstream region.

Moreover, TSS types have a preference for different functional transcript groups defined by the Gene Ontology or tissue expression. For example, GC-rich TSSs correspond to genes

responsible for various binding and protein transport activities. AT-rich TSSs are enriched in processes relating to defense responses to the environment.

To conclude with a very schematic view, in terms of tissue specificity, most tissue-specific genes typically have a TATA box, no CpG island, and often code for extracellular proteins. With the exception of brain-specific genes, CpG islands are found in most of the least tissue-specific genes, which often code for proteins located in the nucleus or mitochondrion. The class of genes with no CpG island nor TATA box are the most common mid-specificity genes and commonly code for proteins located in the membrane (Schug et al., 2005).

3.3 Distal promoter

The distal promoter is located further upstream than the proximal promoter and contains specific TFBSs (Figure 19) that are clustered into enhancer and silencer elements. The enhancers and silencers act either to activate or repress the transcription as well as to increase or decrease the rate of transcription. A typical enhancer is 500 bp long and contains around ten binding sites for at least three different sequence-specific transcription factors, frequently two different activators and one repressor (Levine and Tjian, 2003). Furthermore, boundary/insulator elements prevent the spreading of the activating effects of enhancers or the repressive effects of silencers on heterochromatin (Butler and Kadonaga, 2002). For example, insulator DNAs can prevent an enhancer specific to one gene from inappropriately regulating neighboring genes (Burgess-Beusse et al., 2002).

These regulatory DNAs, enhancers, silencers and insulators are scattered over distances of roughly 10 kb in fruitflies and 100 kb in mammals (Levine and Tjian, 2003). They play an important role in spatial and temporal regulation of gene expression, particularly during development (Howard and Davidson, 2004).

Models of communication between distant protein-DNA complexes include DNA looping (Figure 24), protein tracking or changes in DNA topology, each of which is thought to regulate transcription by increasing the local concentration of factors in the vicinity of the TSS (Barton et al., 1997) (Bagga et al., 2000) (Audit et al., 2002).

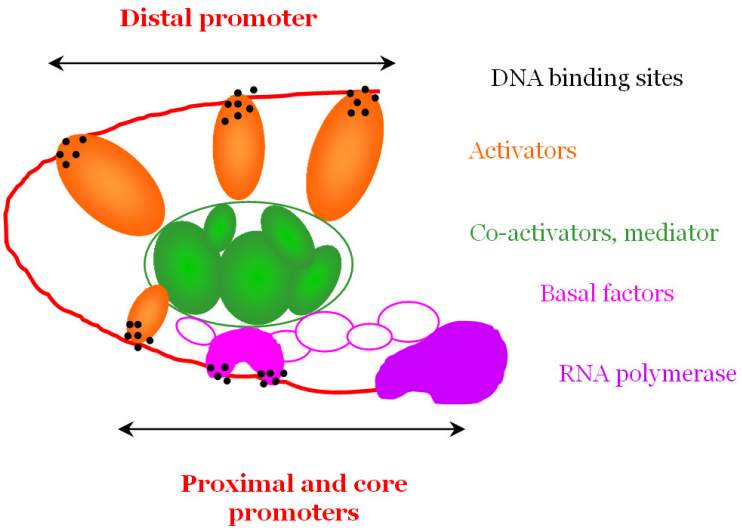


Figure 24. DNA looping mechanism.
This mechanism allows an increased concentration of TFBSs in the vicinity of the TSS. Activators bind to specific DNA control elements in chromatin and interact with multiprotein co-activator machines, such as mediators. The mediator forms a molecular bridge between activators, basal factors and RNA polymerase II. These interactions are possible because DNA is flexible and can form a loop bringing the regulatory regions close together.

RNA Pol II cannot recognize its target promoter directly and cannot initiate transcription without accessory proteins. Instead, this large multisubunit enzyme relies on both general TFs and transcriptional activators and coactivators (Novina and Roy, 1996) (Borukhov and Nudler, 2003) (Figure 24). The core and proximal promoters show specificity both in their interactions with distal promoter elements and with sets of general TFs that control distinct subsets of genes. Thus, the distinct core and proximal promoter regulatory sites, in conjunction with the distal promoter elements such as enhancers, silencers and insulators, define the codes that specify gene expression patterns (Tjian and Maniatis, 1994).

3.4 Recent advances in promoter genomic localization

As previously discussed, TFBSs and CRMs have often been described as being located in the vicinity of the TSS i.e. within the region -1 kb to +200 bp with respect to the TSS (Schmid et al., 2006) (Liu et al., 2003). The recent literature suggests that CRMs can be localized almost anywhere relative to the regulated gene, although they seem to have some localization preferences. The Figure 25 depits and summarizes these different possible locations of the TFBSs on the genomic sequence.

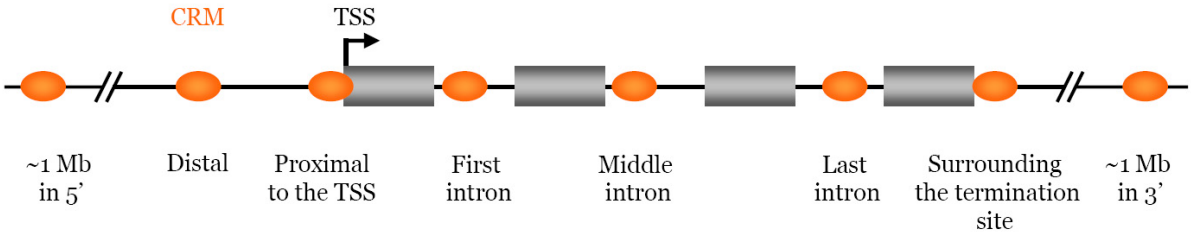


Figure 25. Possible CRMs localization relative to the regulated gene.

In their statistical analysis of over-represented words of 4737 human promoters, Marino-Ramirez highlighted the fact that many of the TFBSs identified showed positional preferences with respect to the TSS (Marino-Ramirez et al., 2004). A reccent genome-wide analysis of predicted CRMs led to the same observation: regulatory modules are preferentially located in specific regions relative to genes (Blanchette et al., 2006).

The CRMs are often located in the non-coding regions of the regulated gene, most often in the upstream flanking region. More recently, several analyses have shown that TFBSs are also sometimes located in a number of other "less conventional" locations such as the 3' UTR regions or even introns (Zhang and Gerstein, 2003) (Cora et al., 2005).

The recent high-throughput study of Blanchette *et al.* offers a comprehensive overview of the favourite positions of CRMs on genes. The results of their study of the position of predicted CRMs with respect to their closest gene are illustrated in Figure 26.

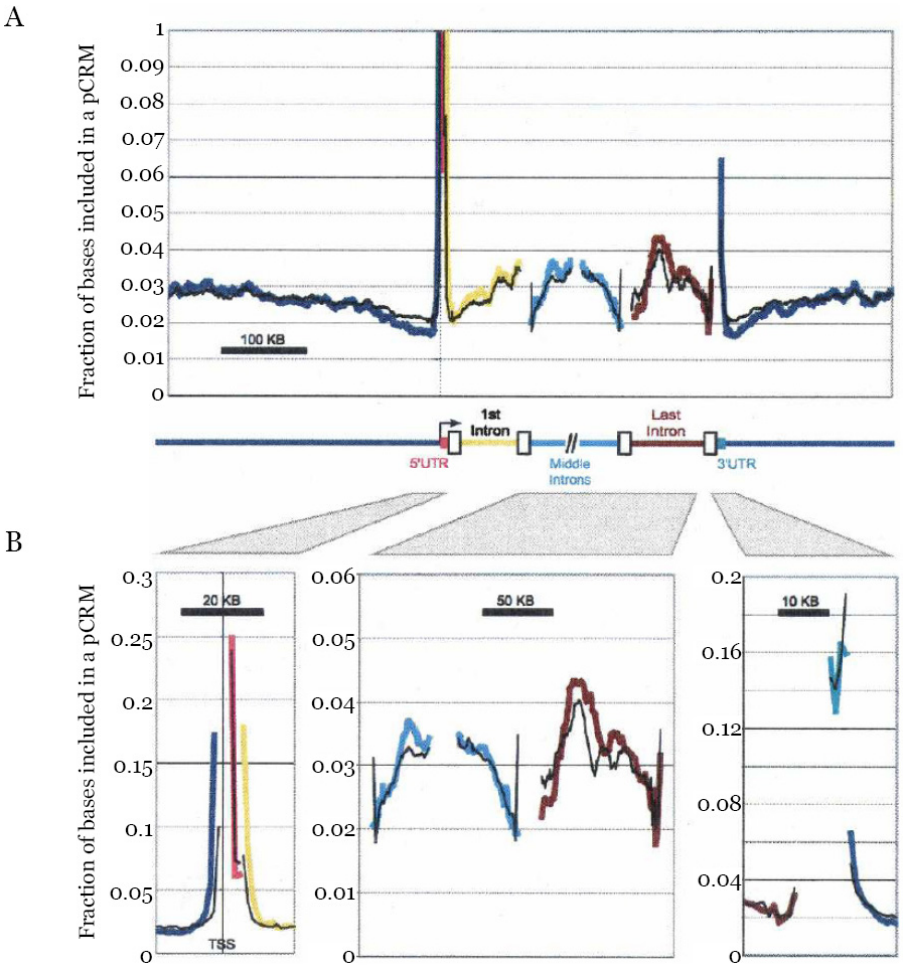


Figure 26. Distribution of predicted CRMs relative to specific regions of genes. Extracted from according to Blanchette *et al.* (Blanchette et al., 2006).

A number of striking observations can be made from these results. First, the regions immediately surrounding TSSs are highly enriched with predicted CRMs (Figure 26). This is explained by the presence of the promoters of the genes in these regions. More surprising is the presence of modules immediately downstream of the TSSs, in the 5'UTR or even also in the first few kilobases of the first intron. These modules may represent alternative TSS but they may also represent a yet underappreciated mode of activation that would take place from downstream proximal binding sites (Blanchette et al., 2006).

Furthermore, the "less conventional" locations of TFBSs that had been suggested before are strongly confirmed in this analysis. Indeed, regions surrounding the site of termination of transcription are also highly enriched with TFBS modules. These modules may represent enhancers that activate the upstream gene via a DNA-looping mechanism (Figure 24). Nevertheless, they may also represent promoters driving noncoding transcripts, antisense relative to the coding gene. Such antisense transcripts may regulate gene expression by a post-transcriptional mechanism (Cawley et al., 2004). Alternatively, these transcripts may

have biological role on their own (cf part 2.2.2.2) and can interfere with sense transcription (Katayama et al., 2005).

These results also surprisingly highlighted the fact that the density of CRMs is the lowest in regions located 10-50 kb upstream of the TSS and, symmetrically, 10-30 kb downstream of the end of transcription. These regions may contain fewer binding sites or binding sites with lower affinity. They may also be depleted in TFBSs. This could be due to structural constraints imposed by the chromatin.

Finally, we observe that the density of predicted CRMs in intronic regions is very low in the close vicinity of exons, except for the first and the last exon. This density then increases with the distance to the closest exon.

Although in simpler organisms, such as yeast, bacteria and viruses, Regulatory Elements (REs) are usually associated with the promoters of their target genes in more complex organisms, especially vertebrates, REs that modulate promoter activity are often positioned remotely from the genes they regulate—sometimes being as far away as a megabase from the transcriptional start site of a gene (Nobrega et al., 2003) (Lettice et al., 2002). They may also be located at distances of several hundred kilobases to over a megabase in either direction from the genes on which they act (Lettice et al., 2003).

3.5 Atypical promoters: the bidirectional promoters

Mammalian genomes contain a class of divergently transcribed genes regulated by putative bidirectional promoters, functional in both directions (Figure 27).

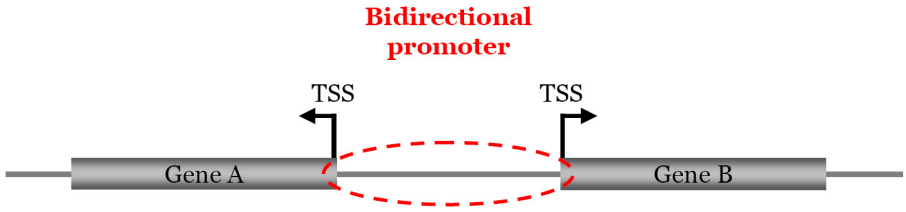


Figure 27. Schematic representation of a bidirectional promoter.

Bidirectional genes were first described in the process of studying a single gene. Adachi and Lieber, 2002 (Adachi and Lieber, 2002), then identified bidirectional genes in a high-throughput study on chromosomes 21 and 22. Since then, a genome-wide analysis of gene organization in the human genomes has been provided by (Trinklein et al., 2004). These computational studies allowed the identification of a major class of gene pairs that are arranged head-to-head on opposite strands with less than 1000 bp separating their TSS. Such

divergently transcribed gene pairs represent more than 10% of all human genes (Trinklein et al., 2004) (Figure 28).

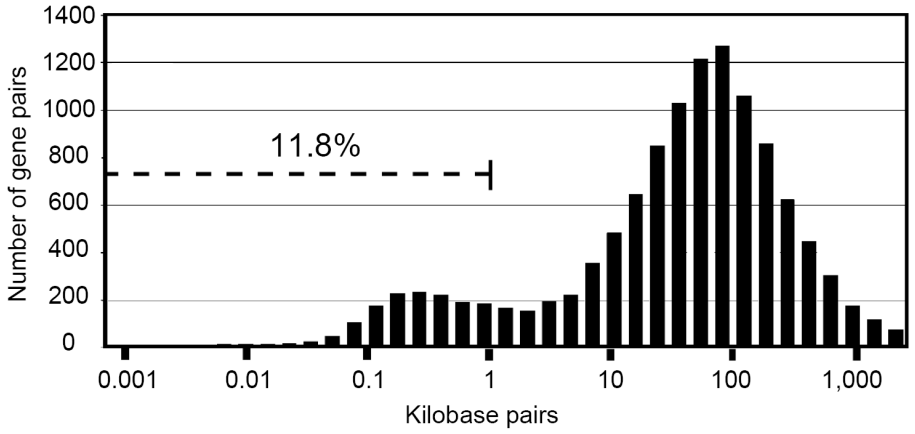


Figure 28. Distribution of distances between 5' ends of genes on opposite strands.
The distribution is bimodal and indicates that 11.8% of genes have TSSs separated by less than 1000 bp. This representation has been extracted from (Trinklein et al., 2004)

Of these 1352 bidirectional gene pairs, 23% have genes whose transcripts are predicted to overlap at the 5'end, whereas 77% are nonoverlapping. Furthermore, the majority (67%) of the nonoverlapping bidirectional promoters are less than 300 bp in length. Nevertheless, no significant correlation between the length of a bidirectional promoter and the degree of expression has been found. For the great majority (81%) of human bidirectionally promoted pairs where the genes have mouse orthologs, the bidirectional arrangement is conserved, suggesting that it is functionally important (Trinklein et al., 2004). In addition, a stronger conservation of distances between bidirectional genes located head-to-head compared to the distances between genes located tail-to-tail has been noticed. This clearly suggests that selection is acting to maintain bidirectional gene organization.

A number of characteristics unique to this class of promoters have been highlighted. In fact, bidirectional promoters have a median GC-content of 66% compared to 53% for nonbidirectional promoters. Furthermore, 77% of the bidirectional promoters are located within a CpG island. Moreover, only 8% of the bidirectional promoters contain a strict TATA-box on either strand, which is not significantly different from what one would expect by chance considering the nucleotide frequencies of the bidirectional promoters. In contrast, 28% of the nonbidirectional promoters contain a TATA-box on the forward strand, which is significantly more than the 18% one would expect by chance. A number of previous studies also observed that bidirectional promoters contain very few TATA-boxes (Qvist et al., 1998) (Dong et al., 2000) (Kawai et al., 2003). All these results suggest that the TATA element regulates the directionality of transcription, but the authors also observed TATA-less

promoters with strong directional activity and some TATA-containing promoters that show activity in both directions. Therefore, the direction of transcription initiation is not always explained by the presence of a TATA-box.

Most bidirectional promoters share at least some regulatory elements that regulate both genes (Trinklein et al., 2004) and individual examples of bidirectional gene pairs have been described, and in some cases it has been shown that the bidirectional promoter indeed regulates the transcription of a gene pair whose levels need to be coordinated expressed for several reasons. For example, some bidirectional promoters serve to maintain a stoichiometric relationship, such as the histone genes (Albig et al., 1997) (Ahn and Gruen, 1999) (Maxson et al., 1983), whereas others regulate the coexpression of genes that function in the same biological pathway (Schmidt et al., 1993) (Sugimoto et al., 1994) (Momota et al., 1998), or control expression through different time points such as genes involved in the cell cycle (Guarguaglini et al., 1997). Other bidirectional promoters provide coordinated responses to induction signals such as heat shock (Hansen et al., 2003). DNA-repair genes are overrepresented in the bidirectional class in the human genomes (Adachi and Lieber, 2002) (Trinklein et al., 2004). This is also the case for genes encoding chaperone proteins, mitochondrial genes as well as a class of DEAD-box RNA helicases.

If the majority of the promoter segments between two bidirectional genes initiate transcription in both directions, for a minority, the bidirectional promoters induces the transcription initiation of one gene while inhibiting transcription in the other gene (Trinklein et al., 2004). The directionality of promoter activity may also be regulated to some degree in a cell type-specific manner. All these observations are in agreement with other studies that found that genes sharing a bidirectional promoter have a higher probability of being coordinately expressed than random pairs of genes (Engstrom et al., 2006).

In fact, most bidirectional promoters act as inseparable units that may provide a unique mechanism of coordinated or anti-coordinated regulation of transcription for a significant number of mammalian genes (Trinklein et al., 2004).

3.6 TFBS and promoter conservation

3.6.1 TFBS conservation

The rate of conservation of TFBSs varies according to the binding TF (Sauer et al., 2006). For example, human and mouse sequence conservation of the binding sites for MyoD, MEF-2, SRF or NF-AT1 are very high (between 96.7 and 100%) while TFBSs for Sp1, C/EBPα, AP-

2αA or GATA-1 exhibit a lower sequence conservation rate (between 44 and 63.7%) (Sauer et al., 2006) (Wasserman et al., 2000). The TFs recognizing highly conserved TFBSs may have essential roles in gene regulation for both human and rodents which increase the selective pressure on their TFBSs.

There is experimental evidence that at least some of the factors with a lower conservation rate (Sp1, GATA-1) have clustered TFBSs (Hardison et al., 1993) (Hermfisse et al., 1996). As examples, TRANSFAC documents four AP-2αA TFBSs between -230 and -110 bp upstream of the TSS of the human MT2A gene and six c-Myb TFBSs in the human SFRS2 gene (-366 to +16 bp relative to the TSS) (Sauer et al., 2006).

The conservation rate for TFBSs also differs depending on the "biological processes" in which the regulated genes are involved (Sauer et al., 2006) (Table 5).

GO slim term	*N*	C_{seq} (%)	p_{seq}
Transcription regulator activity	262	80.5	3.1E−49
Hydrolase activity	199	79.4	2.1E−35
Cell death	195	77.9	2.4E−32
Regulation of biological process	745	77.3	5E−120
Nucleic acid binding	356	77.2	3.8E−57
Catabolism	187	77.0	1.4E−29
Signal transducer activity	714	75.8	2E−106
Development	526	75.1	9.5E−76
Protein binding	779	74.8	1E−110
Response to stimulus	703	74.8	8E−100
Cell communication	602	74.1	3.1E−82
Enzyme regulator activity	112	72.3	9.1E−14
Metabolism	1197	71.1	9E−139
Catalytic activity	646	67.5	1.5E−60
Kinase activity	136	66.9	3.6E−12
Cell motility	64	65.6	4.5E−05
Electron transport	95	65.3	1.6E−07
Transport	388	62.1	2.8E−25
Transferase activity	223	61.9	3.2E−14
Oxidoreductase activity	187	59.9	3.6E−10
Transporter activity	350	59.7	6.2E−19
Electron transporter activity	62	56.5	0.03223

Table 5. Gene function dependency of the TFBS conservation rate.
N represents the number of TFBSs linked to the GO slim term. The conservation rate C_{seq} for human-rodent comparisons is defined as the fraction of all N TFBSs that are considered as conserved. The *p-value* p_{seq} to obtain the observed (or a bigger) difference between the conservation rate and the background conservation rate by chance estimates the significance of the conservation. This table is extracted from the study of Sauer *et al.* (Sauer et al., 2006).

TFBS regulating genes with activities such as *transcription regulation* (C_{seq} of 80.5%), *cell death* (77.9%), *regulation of biological processes* (77.3%), *nucleic acid binding* (77.2%), *signal transduction* (75.8%) or *development* (75.1%) have high conservation rates.

Highly conserved regulatory sequences are found to cluster in the vicinity of genes implicated in transcriptional regulation and in early development (Bejerano et al., 2004) (Sandelin et al., 2004b) (Woolfe et al., 2005) and the majority of those tested in vivo (5/7 in mice, (Nobrega et al., 2003); 23/25 in fish, (Woolfe et al., 2005)) drive expression of reporter genes in a temporal and spatial specific manner during early development (McEwen et al., 2006). Many other studies around specific developmental genes have also identified highly conserved noncoding sequences between humans and fish that have enhancer activity (Zerucha et al., 2000) (Barton et al., 2001) (Blader et al., 2003) (Dickmeis et al., 2004) (Goode et al., 2005). The association of these highly conserved sequences to genes implicated in the regulation in early development is most likely a result of the fundamental nature of the developmental process in vertebrates.

On the other hand, TFBS regulating genes involved in *transport* (62.1%) and having certain catalytic activities, e.g. *kinase* (66.9%), *transferase* (61.9%) and *oxidoreductase activity* (59.9%), exhibit low conservation rates (Table 5).

Fish–mammal genomic comparisons have proved powerful in identifying conserved noncoding elements likely to be *cis*-regulatory in nature, and the majority of those tested *in vivo* have been shown to act as tissue-specific enhancers associated with genes involved in transcriptional regulation of development (McEwen et al., 2006). Although most of these elements share little sequence identity to each other, a small number are remarkably similar and appear to be the product of duplication events. Being the most evolutionary distant extant vertebrates for which whole genome information is available, mammals and fish comparison provide high stringency for the detection of vertebrate-specific regulatory elements. These elements with high conservation have remained practically unchanged in the 450 million years since the divergence of fish and mammals.

3.6.2 Promoter conservation

Functional regions such as protein-coding exons are known to maintain sequence similarity across large evolutionary distances, such as that between human and pufferfish *Takifugu rubripes* (Aparicio et al., 2002). The tendency of functional regions to be conserved under selection pressure during evolution explains the importance of studying the non-coding region conservation during evolution. This approach allows the identification of the promoter regions that have been conserved compared to non-coding regions that have been altered by random genome dynamics. As a consequence, the conserved promoter regions are considered to be potentially functional. In contrast, intronic and intergenic conserved

elements "evaporate" much more rapidly, limiting the sensitivity of distant noncoding sequence comparisons.

Although only about 1.2% of the human genome appears to code for proteins (Lander et al., 2001) (Venter et al., 2001), it has been estimated that as much as 5% is more conserved than would be expected from neutral evolution since the split with rodents (Waterston et al., 2002) (Chiaromonte et al., 2003). Several studies have found significant numbers of specific noncoding segments in the human genome that appear to be under selection, using a threshold for conservation of 70 or 80% identity in multiple species over more than 100 base pairs (bp) (Loots et al., 2000) (Dermitzakis et al., 2002) (Dermitzakis et al., 2003).

To quantify the decay of noncoding sequence conservation with evolutionary distance, Prabhakar *et al.* analysed whole-genome conserved noncoding sequences (CNS) sets using three-way genome alignments: human–mouse–rat (HMR), human–mouse–chicken (HMG), human–mouse–frog (HMX), and human–mouse–fish (HMF) (Prabhakar et al., 2006). As expected, the more distantly related genomes exhibit markedly less conservation relative to human. The CNSs have been classified into four categories (Figure 29).

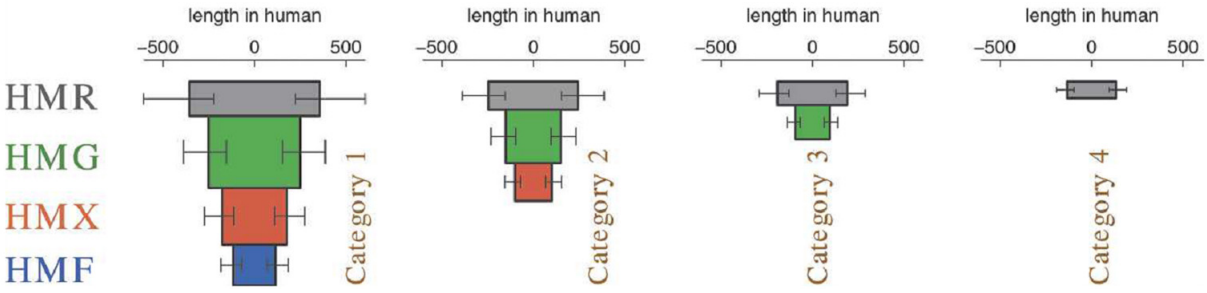


Figure 29. Whole-genome noncoding conservation and the "funnel principle": correlation between ancient and recent noncoding conservation.
Blocks represent the CNSs. Block width is proportional to median CNS length in human, and block area is proportional to the statistical significance of the alignment between the genomes. Block height thus represents degree of evolutionary constraint at the basepair level. Error bars mark the range from the 25th to the 75th percentile of CNS length. The "funnel principle" takes its name from the funnel-like shape of ESPs. This figure has been extracted from (Prabhakar et al., 2006).

Category 1 consists of human–rodent CNSs that overlap human–mouse–chicken, human–mouse–frog, and human– mouse–fish CNSs. Category 2 extends only to human– mouse–frog, Category 3 to human– mouse–chicken, and Category 4 is restricted to human–rodent alone. Interestingly, this study highlighted the fact that the longer and more constrained a mammalian noncoding sequence, the greater its evolutionary conservation in distant vertebrates and vice versa (Prabhakar et al., 2006).

Similarly, a whole-genome comparison between human and the pufferfish, *Fugu rubripes*, identified nearly 1,400 highly conserved non-coding sequences (Woolfe et al., 2005). In another study, despite the fact that only about 4% of the human genome can be reliably aligned to the chicken genome (at an average of 62.9% identity where an alignment is found), and less than 1.8% of the human genome aligns to *fugu* (with an average of 60% identity) (Bejerano et al., 2004), some conserved non-coding sequences exhibit extremely high levels of conservation with orthologous regions in the chicken genome [467 out of 481 (467/481) = 97% of the elements having an average of 95.7% identity, 29 with 100% identity] and in the *fugu* genome (324/481 = 67.3% of the elements having an average of 76.8% identity),. Given the extreme evolutionary divergence between these species, it is likely that these sequences, which are never found in invertebrate genomes, are essential to all vertebrates. These highly conserved non-coding sequences, frequently located in and around genes, are likely to form part of the genomic circuitry that uniquely defines vertebrate development (Bejerano et al., 2004) (Woolfe et al., 2005).

Various genomic level conservation analyses oriented through the TSS regions have also been performed. For example, an analysis of genomic regions of -700 to +300 bp with respect to the TSS in human, mouse and rat, highlighted an average conservation of 55% between human-rodent homologous promoter pairs and 85% between mouse-rat homologous promoter pairs (Xuan et al., 2005). On the other hand, the average sequence identity in the "upstream regions" of human and mouse genes has been reported to be approximately 70-75% (Waterston et al., 2002), which is significantly higher than previous estimates (45% for (Suzuki et al., 2004b) and 55% for (Xuan et al., 2005)). This may have been due to the fact that Waterston *et al.* used upstream 200 bp regions compared to -1000 bp to +200 bp and -700 to +300 bp with respect to the TSS respectively. These results suggest the average sequence identity is the highest in the close upstream vicinity of the TSSs (Waterston et al., 2002). A previous report also indicated that the number of alignable sequences decreases relatively rapidly in the upstream regions of the TSSs (Jareborg et al., 1999).

A recent study of larger regions (4kb centered on TSS) also revealed that 51% of the bases in promoters (44% upstream and 58% downstream of the TSS) could be aligned across human, mouse, rat and dog genomes (Xie et al., 2005). These proportions are much higher than for intronic regions (34%) or for the whole genome (28%). Such conservation in promoters reflects the presence of important elements under evolutionary selection pressure.

As previously discussed, promoter sequences are frequently conserved in a block manner (Suzuki et al., 2004a). This is inconsistent with the initial descriptions of the sequence similarity among promoters, which indicated that the independent alterations of the

nucleotides are distributed in a gradually increasing manner in direct proportion to the distance from the TSSs. Although the results of a previous study using 41 human-mouse promoter pairs suggested the block structure of the sequence conservation in the promoters, it was considered likely to be an artifact of the alignment program used (Jareborg et al., 1999). Nevertheless, in at least one-third of the promoters, such a conservation block is present in the 1 kb region upstream of the TSS. Within these blocks, the sequence similarity is relatively uniform, designing flat conservation patterns, regardless of their length and with an average identity of 65%. Using relics of ancestral repeats the overall sequence similarity between human and mouse at neutral sites has been estimated to be 53-54% (Waterston et al., 2002). If the regional variations of the neutral substitution rate are ignored (Hardison et al., 2003), the sequence identity is approximately 10% higher in the sites within the promoter blocks. This difference implies that some parts of the promoters are subjected to selective pressure. The relatively flat patterns of sequence similarities within blocks may result from the fact that the positions of the TFBSs are different between distinct genes, allowing degeneracy within them to some extent. It is also possible that additional sequences other than the direct TFBSs themselves might also be conserved, considering that the cognate sequences of the TFs are typically 6-10 bp long (Wray et al., 2003). Thus, particular subregions of the promoter may not have been allowed to undergo free sequence divergence because the overall base composition or relative positions of TFBSs needed to be preserved (Suzuki et al., 2004a).

This block structure clearly seems to have been formed as a consequence of a selective pressure since alterations occurring inside the transcriptional regulatory modules in the promoters would mostly be unfavorable for proper biological functions (Suzuki et al., 2004a). Most of the previously characterized TFBSs are located within the blocks of conservation. For these TFBSs, the cognate sequences as well as the relative positions of the TFBSs and distances to the TSSs are preserved. Alterations that occurred outside blocks may generally have been tolerated since blocks seemed completely lost from the corresponding parts of the other genomes.

Some blocks alterations might have led to the acquisition of altered modes of transcriptional modulation. Phenotypic effects of such modulation are unclear since it has been reported that polymorphisms that cause an approximately twofold difference in transcription activation activities frequently occur without showing organismal phenotypes within human populations (Rockman and Wray, 2002). Repetitive elements at the boundaries of the blocks are thought to contribute to such transcriptional modifications and indeed, repetitive elements are found at the boundaries of ~46% of the blocks in either the human or mouse

genomes (Suzuki et al., 2004a). There are a number of examples in which retroelements integrated in the vicinity of TSSs became involved in transcriptional regulation via changes in their sequences (Norris et al., 1995) (Vansant and Reynolds, 1995) (Hamdi et al., 2000). It is likely that such variations have accumulated during evolution and have formed the genetic background to drive speciation during certain periods of time.

The length of the conservation blocks is shorter in the promoters of brain-specific genes and of TFs coding genes, which suggests that alterations within the proximal regions of the TSSs have been accumulated for these gene populations and that the evolutionary diversifications in the transcriptional modulations should be the most marked for these genes (Suzuki et al., 2004a).

Several studies have discovered a correlation between the locations of highly conserved sequence elements and genes involved in specialized biological processes, and notably the development and the regulation of transcription (Bejerano et al., 2004) (Sandelin et al., 2004b) (Woolfe et al., 2005). These highly conserved sequences probably exist in all vertebrate species and are essential for vertebrate development. Iwama and Gojobori focused in their study on the conservation of sequences directly upstream of 3055 orthologous human–mouse gene pairs (Figure 30) (Iwama and Gojobori, 2004).

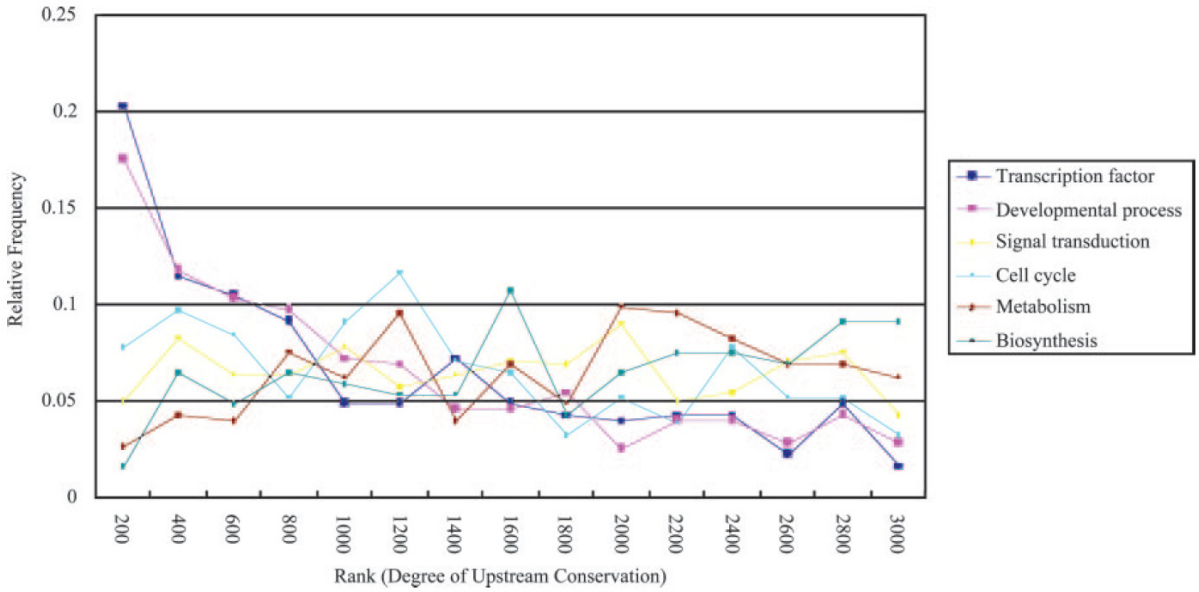


Figure 30. Relative frequency of genes annotated with six functional categories with respect to the degree of upstream sequence conservation.
The relative frequency for the six functional categories is shown, along with ranks that represent the degree of upstream sequence conservation. For instance, rank 200 includes the top 1–200 upstream-conserved genes, and the points on each line graph represent the relative frequency within the corresponding rank for the corresponding functional category. The genes under analysis were subdivided into six functional categories according to the Gene Ontology (GO) annotation. Extracted from (Iwama and Gojobori, 2004).

The authors have shown that transcription factor genes and developmental process-related genes show the highest degree of upstream sequence conservation, while genes involved in metabolism and cell cycle show less upstream sequence conservation. To complete this study, they further focused on the TF genes involved in developmental processes (Figure 31).

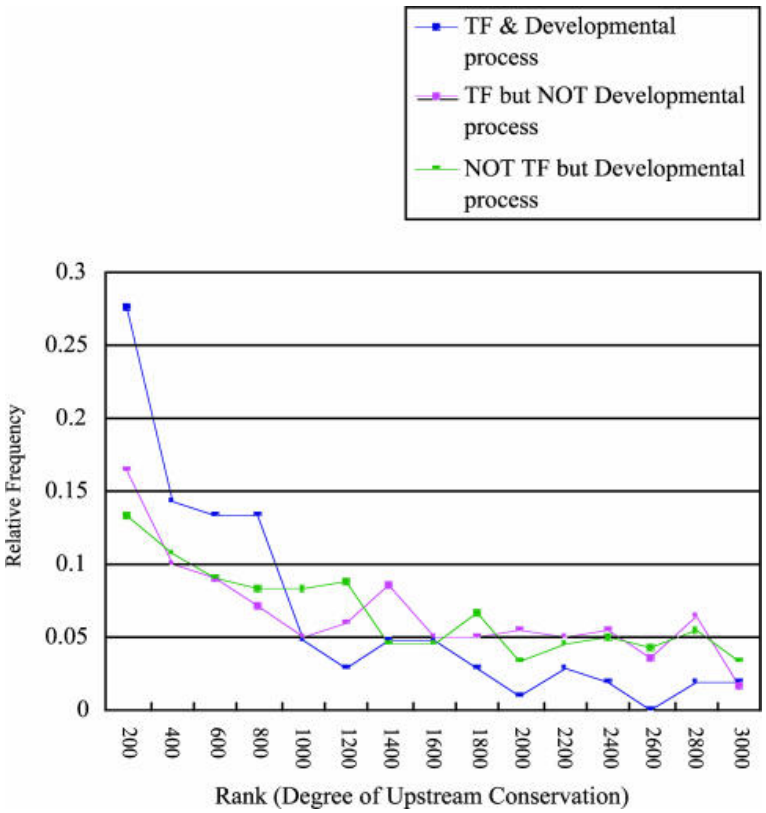


Figure 31. Relative frequency of genes annotated with transcription factor and/or developmental processes with respect to the degree of upstream sequence conservation.
The relative frequency for developmental process-related TF genes is shown, along with ranks that represent the degree of upstream sequence conservation, together with the relative frequencies of TF genes that are not involved in the developmental process and the developmental processrelated non-TF genes. Extracted from (Iwama and Gojobori, 2004).

The studied developmental process-related TF genes displayed an even higher degree of upstream sequence conservation, with roughly 27% of the genes appeared in the top 200 upstream-conserved rank (Iwama and Gojobori, 2004).

Finally, it should be stressed that conservation rates in promoters is not a uniform property since an acceleration of primate promoter evolution has also been reported (Taylor et al., 2006). It appears that evolution within core promoters has been relatively rapid for perhaps 25 million years of primate evolution and that this may be a distinctive characteristic of our order. This clearly highlights that the understanding between recent and ancient sequence conservation is likely to grow even stronger when more mammalian genomes, and notably

primates or dog, will be added to the classical highly studied human–mouse–rat trio (Margulies et al., 2005).

Chapter 4 - *In silico* promoter analysis

In order to understand the transcriptional network of human genes, it is essential to characterize their regulatory regions, which include the promoters. To this end, one of the challenges confronted by both experimental and bioinformatics researchers has been the identification of what kind of functional sequence elements reside in which parts of the promoters and how they serve as modulators of transcription.

In the post-genomic era, the availability of high-throughput experimental data and complete genome sequences has been greatly favorable to high-throughput promoter sequence analysis using bioinformatics tools. Such bioinformatic approaches can contribute to the identification of promoter regions of potential interest and of *cis*-regulatory elements before engaging in time-consuming biochemical characterization.

In this chapter, we will present some general principles and bioinformatics approaches dedicated to *in silico* promoter analysis. Fisrt, we will focus on approaches taking into account the promoter during evolution. Second, we will present methods used to find the transcriptional regulators common to a set of genes that are involved in the same biological pathway, in the same cellular process, in response to the same stimulus, or in the same disease.

4.1 Appropriate genomic sequence size

The first problem in promoter analysis is the determination and the localization of the promoter sequence to be analyzed. Most current promoter analyses first focus on the proximal promoter region. Thus, in such *in silico* analyses, knowledge of the precise Transcription Start Site (TSS) is of critical importance.

Many mRNA described in databases are incomplete at their 5′ ends and the estimation of their TSS can be more or less inaccurate particularly in the case of unidentified first exon/intron pairs (Yamashita et al., 2005). This implies that the proximal genomic region upstream of the potential TSS determined by a genomic localization of an mRNA sequence is sometimes inappropriate. Furthermore, the prevalence of promoters having a general broad distribution of TSS in mammals implies that one cannot consider the most extreme 5′ end of

the longest cDNA as a true full-length transcript (Carninci et al., 2006). Thus, promoter analysis can be problematical depending on the structure of the studied gene.

Bioinformatics promoter studies presented in the literature use a wide range of different positions related to the TSS to determine the promoter sequences to be analyzed. For example, Marino-ramirez *et al.* define their putative promoter regions as the genomic sequences running from -2,000 to +1,000 bp (Marino-Ramirez et al., 2004). Xie *et al.* studied the non-coding sequence contained within a 4 kb window centered on the TSS (Xie et al., 2005), while Suzuki *et al.* considered the 1 kb proximal regions (Suzuki et al., 2004a) and Di Cara *et al.* focused on the core promoter defined as 500 bp upstream of the TSS (Di Cara et al., 2005).

In simple organisms such as yeast or worm, the intergenic sequences are usually very short. Therefore, previous studies were able to obtain meaningful results or make reliable predictions using 500–600 bp as the putative promoter length (Hughes et al., 2000) (GuhaThakurta et al., 2002). In mammals, the intergenic sequences may be very long and the regulatory sequences may be located at large distances from the TSSs. Approximately 40% of human genes have completely non-coding first exons. For most of these genes, the promoter region occurs well upstream of the translation start codon (ATG) because the first intron tends to be longer than average (Davuluri et al., 2001). In higher eukaryotes, regulatory elements have also been found in the first intron (Mathew et al., 2004). Based on this knowledge, Chang *et al.* analyzed much larger promoter sequences. Indeed, they defined the sequence range that is mostly likely to contain regulatory signals as 10 kb upstream and 5 kb downstream of the TSS (Chang et al., 2006).

Furthermore, Brown and Feder demonstrated that in some cases different expression profiles of the same gene in several *Drosophila melanogaster* strains can be observed although the proximal promoter sequences are identical (Brown and Feder, 2005). Hypothetically, genes differing in expression should have more proximal promoter polymorphisms than those whose expression is conserved. Nevertheless, the analysis of the proximal promoters of six genes with significantly different mRNA expression in *Drosophila* revealed that they were identical in sequence. This highlights the fact that regulatory sequences external to the proximal promoter, such as distal promoter (enhancers…) must play a greater role than the proximal promoter in the observed transcriptomic variations.

In promoter analysis, each new promoter constitutes a unique case with novel properties, surprises, challenges and difficulties and thus, bioinformatics approaches need to be adapted to each particular study.

4.2 Promoter and TSS prediction/validation

The variable and sometimes large 5′ UTR of mammalian messengers make any reference to the start of translation rather weak. The poor sequence conservation between start of transcription and start of translation (Werner, 2002) makes computer prediction a complex and sometimes unreliable task.

Two ways are available to localize the TSS position on a genomic sequence. Indeed, promoter experimental databases or promoter prediction programs can be used for this task.

4.2.1 Experimental databases

In order to overcome the problem of obtaining precise information of 5′ end termini based on cDNA sequences, promoter databases for mammals have been constructed by collecting experimentally identified TSSs.

4.2.1.1 DBTSS, DataBase of Transcriptional Start Sites

DBTSS was first constructed in 2002 based on precise, experimentally determined 5′ end clones (Suzuki et al., 2002), obtained using the "oligo-capping" method.

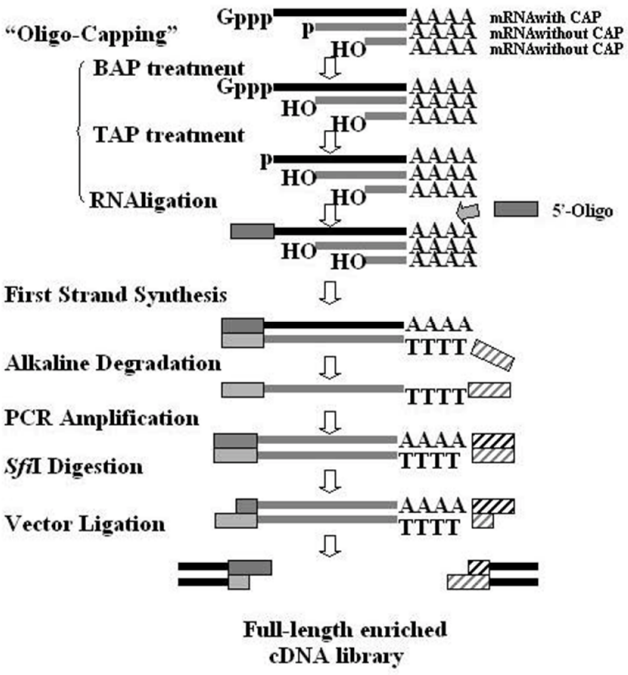


Figure 32. "Oligo-capping" method
The cap structure of the mRNA was replaced with the 5'-oligo by the "Oligo-capping" method, which consists of three enzymatic reaction steps. 1: bacterial alkaline phosphatase (BAP) hydrolyses the phosphate of truncated mRNA 5'-ends whose cap structures have been broken down. 2: tobacco acid pyrophosphatase (TAP) removes the cap structure, leaving the phosphate at the 5'-end. 3: T4 RNA ligase, which requires a phosphate at the 5'-end as its substrate, selectively ligates the 5'-oligo to the 5'-end that originally had the cap structure. Using "Oligo-capped" mRNA, first strand cDNA was synthesized with dT adapter primer. After alkaline degradation of the RNA, first strand cDNA was amplified by PCR, digested with restriction enzyme SfiI and cloned into a plasmid vector. RNA and DNA molecules are represented by solid lines, the 5'-oligo by gray boxes, and PCR primers by shaded boxes. Gppp: cap structure; p: phosphate; OH: hydroxyl.

Other human cDNA libraries enriched in clones containing the cap structure have been constructed to systematically explore the 5'end structure of expressed genes. Each sequence has been mapped to the human genome sequence to identify its TSS.

In the first published version of the database (Suzuki et al., 2002), 111,382 clones match both RefSeq mRNA (7,889 Refseq entries) and the genome sequence corresponding to genes. They observed that 34% of the RefSeq sequences could be extended toward the 5' end. On average, the data extended the RefSeq sequences by 87 bases. In 2004, the number of clones had increased to 190,964, covering 11,234 genes (Suzuki et al., 2004b). Today, the number of human clones has drastically increased to 1,359,000 (Yamashita et al., 2006) (Table 6).

	No. of clones	No. of genes/no. of RefSeq	No. of TSSs	No. of promoters
Human	1 359 000	15 262/19 753	452 117	30 964
Mouse	364 487	14 162/14 746	149 876	19 023
Zebrafish	32 263	3061/3075	15 198	3382
Malaria	10 236	1527/-	6908	-
Schyzon	22 923	3635/-	14 029	-

Table 6. Statistics of DBTSS.
Extracted from (Yamashita et al., 2006).

The 1,359,000 clones in DBTSS correspond to 19,753 RefSeq cDNAs (Table 6). Since RefSeq cDNAs contain splicing variants as separate entries, clustering of clone information was performed depending on their coordinate in the genome sequence; if their sequences overlapped, they were considered to be the same locus. After clustering, the data corresponded to 15,262 genes (Table 6). Furthermore, information about potential alternative promoters is newly available in the current version because the number of 5' end clones is now sufficient to determine several promoters for one gene. To determine alternative promoters, all the TSSs from the same locus were considered. The clones were clustered using a 500 base interval, and each cluster was defined as a promoter. According to this procedure, 6,954 human loci and 9,886 mouse loci have only one promoter while 8,308 human loci and 4,276 mouse loci have two or more promoters. In addition, TSS information for 3,061 zebrafish (*Danio rerio*) genes, 1,527 malaria (*Plasmodium falciparum*) genes and 3,635 schyzon (*Cyanidioscyzon merolae*) genes (a red algae model organism) were also included in the latest version. In the future, DBTSS will be updated with new TSS information from organisms such as macaque, when the data becomes publicly available.

The DBTSS promoter sequences are provided in the database as genomic sequences from -1,000:+200 with respect to the experimentally determined TSS. DBTSS is accessible at http://dbtss.hgc.jp.

4.2.1.2 EPD, Eukaryotic Promoter Database

The Eukaryotic Promoter database (EPD) is an annotated non-redundant collection of eukaryotic pol II promoters for which the TSS has been determined experimentally (Cavin Perier et al., 1998). The underlying definition of a promoter is that of a transcription initiation site. All information in EPD is derived from independent examination and interpretation of experimental data reported in cited research publications. Many TSS described in the literature have not been included in EPD because the underlying experimental evidence did not meet the minimal requirement for inclusion. In the current version (Schmid et al., 2006), access to promoter sequences is provided by pointers to positions in the corresponding

genomes. Promoter evidence comes from conventional TSS mapping experiments for individual genes, or, starting from release 73, from genome annotation projects.

EPD has undergone drastic changes since the beginning of the functional genomics era. Before, promoters were mapped for one gene at a time by techniques such as nuclease protection assay and primer extension analysis. The corresponding EPD entries were the result of a critical examination and independent interpretation of data published in paper-based journal articles. Today, TSSs are mapped for a whole genome with high-throughput technologies such as 5'SAGE (Hashimoto et al., 2004) or CAGE (Carninci et al., 2005). The resulting data are disseminated in machine-readable form over the internet. As a consequence, EPD entries are now largely generated by intelligent Perl scripts with built-in quality control procedures rather than by critical readers of scientific articles. An overview of publicly available genome annotation data useful for promoter mapping is given in Table 7.

5' EST sequences from oligo-capped cDNA libraries			
Human	http://dbtss.hgc.jp/	400 225	Suzuki *et al.* (Suzuki et al., 2004b)
Mouse	http://dbtss.hgc.jp/	580 209	Suzuki *et al.* (Suzuki et al., 2004b)
Drosophila	Sequences available from Genbank/EMBL, accession numbers extractable from Unigene (23), Unilib IDs 23941 or 23942	102 617	Stapleton *et al.* (Stapleton et al., 2002)
Arabidopsis	ftp://pfgweb.gsc.riken.jp/rafl/	92 654	Seki *et al.* (Seki et al., 2002)
5' sequence tags (5'SAGE, CAGE, GIS ditag)			
Human	http://5sage.gi.k.u-tokyo.ac.jp/	22 546	Hashimoto *et al.* (Hashimoto et al., 2004)
Human	http://fantom31p.gsc.riken.jp/cage/download/hg17/	5 992 395	Carninci *et al.* (Carninci et al., 2005)
Mouse	http://fantom31p.gsc.riken.jp/cage/download/mm5/	11 567 973	Carninci *et al.* (Carninci et al., 2005)
Mouse	ftp://fantom.gsc.riken.jp/FANTOM3/GIS/	225 914	Ng *et al.* (Ng et al., 2005)
Reference sequence collections from oligo-capped cDNA libraries			
Rice	ftp://cdna01.dna.affrc.go.jp/pub/data/CURRENT	30 598	Kikuchi *et al.* (Kikuchi et al., 2003)

Table 7. Summary of currently accessible mass genome annotation data for promoter mapping.

The EPD sequences are given from -9,999 to +6,000 relative to TSS. EPD aims to achieve promoter coverage for three model organisms (human, *D. melanogaster* and rice) as soon as possible. EPD can be accessed at http://www.epd.isb-sib.ch .

Other specialized promoter databases exist such as HemoPDB (Pohar et al., 2004) that specializes in promoters of genes of the hematopoietic system. Nevertheless, notwithstanding the DBTSS, EPD and other dedicated databases, promoter data are still very limited and in certain cases, prediction programs represent a good alternative.

4.2.2 *Ab initio* prediction programs

A number of promoter prediction programs based on different concepts have been developed over the last ten years. The underlying principle is that properties of promoter regions are different from properties of other genomic DNA. The Table 8 describes some promoter prediction programs.

Program	Reference	Concept	Technology
GRAIL	(Matis et al., 1996)	TATA-, GC-, CAAT-boxes, the INR Distance contraints between these elements	Neural Network
Promoter 2.0	(Knudsen, 1999)	TATA-box motif CpG island	Artificial Neural Networks
PromoterInspector	(Scherf et al., 2000)	Libraries of IUPAc words described in the promoter regions	Unsupervised learning approach
FirstEF	(Davuluri et al., 2001)	CpG island	Linear and quadratic discriminant analyses
McPromoter	(Ohler et al., 2002)	TATA-box, INR, DPE CpG island	Interpolated Markov model
Dragon PF	(Bajic et al., 2002)	CpG island	Artificial Neural Networks
CpGProD	(Ponger and Mouchiroud, 2002)	CpG island	CpG frequency followed by statistics calculations
Eponine	(Down and Hubbard, 2002)	CpG island TATA-box motif	Relevance Vector Machine
PromH	(Solovyev and Shahmuradov, 2003)	TATA-box motif Homology with orthologous promoters	Linear and quadratic discriminant analyses
Dragon Gene Start Finder (GSF)	(Bajic and Seah, 2003)	CpG island	Artificial Neural Networks

Table 8. Promoter prediction programs.

This table illustrates that many approaches have been used, such as the presence of CpG islands close to the TSS locations, the presence of specific TFBSs such as the TATA-box or the INR motifs, statistical properties of proximal and core promoters as opposed to other genomic sequences.

Recognition technologies employed in promoter prediction programs are based on neural networks, linear and quadratic discriminant analyses, unsupervised learning, Relevance

Vector Machine, statistical properties of promoter regions, interpolated Markov model, or a combination of these.

For many promoter prediction programs, authors report good performance on human chromosome 22, but this is the second most G+C rich human chromosome. Recently, Bajic *et al.* (Bajic et al., 2004a) tested these programs on the whole human genome in order to reduce the bias introduced by the different genomic test sets. This study highlighted the fact that no program can predict non-CpG-island-related promoters satisfactorily. DragonGSF and Eponine give something more than one false-positive prediction for every two true-positive predictions. CpGProD and Eponine have a great accuracy in predicting TSSs; nevertheless CpGProD is exclusively restricted to CpG-island-related promoters while Eponine is restricted to G+C-rich promoters containing a TATA-box. The best general purpose promoter prediction programs appear to be DragonGSF (which has a preference for CpG-island-related promoters) and FirstEF. DragonPF and FirstEF predict the most diverse sets of promoters. Very good performance is obtained by McPromoter, but it is very slow. Promoter2.0 produces predictions close to, or worse than random guessing (Bajic et al., 2004b).

Large scale computational methods to identify promoters have shown that approximately only 50% of promoters could be correctly predicted from the genomic sequence (Bajic et al., 2003). On the other hand, it should be noted that theses methods suffer from a high false-positive rate or only predict a very specific subset of core promoter, e.g. promoters linked to CpG islands (Down and Hubbard, 2002).

The most efficient solutions currently available use CpG islands, first exon properties, TATA-box and several other properties from the core promoter region. Nevertheless, these methods do not possess sufficient accuracy for the positional of TSS predictions. This suggests that, although these features are important, they are not sufficient for accurate TSS location by computational means. Thus, present algorithms aimed at TSS prediction have proven unsatisfactory (Bajic et al., 2004b), especially for detection of non-CpG-island-related promoters. Furthermore, current promoter prediction programs are not able to precisely determine the TSS and thus cannot effectively detect alternative TSSs.

4.3 TFBSs predictions on the genomic sequence

Transcription factors can tolerate widely varying target sequences, resulting in computational binding profiles of low specificity. Such weak patterns become impossible to

distinguish when regulatory regions are embedded within long candidate regions (Thompson et al., 2004b).

A large number of transcription factors (TFs) have been identified and their sequence-specific binding to promoter elements has been shown to play a central role in regulation (Mitchell and Tjian, 1989) (Novina and Roy, 1996). As many of the TFBSs are short (5-25 bp) (Qiu, 2003) and degenerated sequences, a huge quantity of TFBSs are predicted on genomic sequences. Therefore, it is a difficult problem to discriminate the genuine TFBSs, which have biological significance *in vivo*, from other sequences that occur randomly and frequently in the large volume of human genomic sequences (Fickett and Wasserman, 2000).

Several methods are currently used to define TFBSs, they are presented in the following session.

4.3.1 TFBSs representation

The first step towards building models for predicting TFBS involves data collection. To illustrate the process, the example of MEF2 is shown in Figure 33.

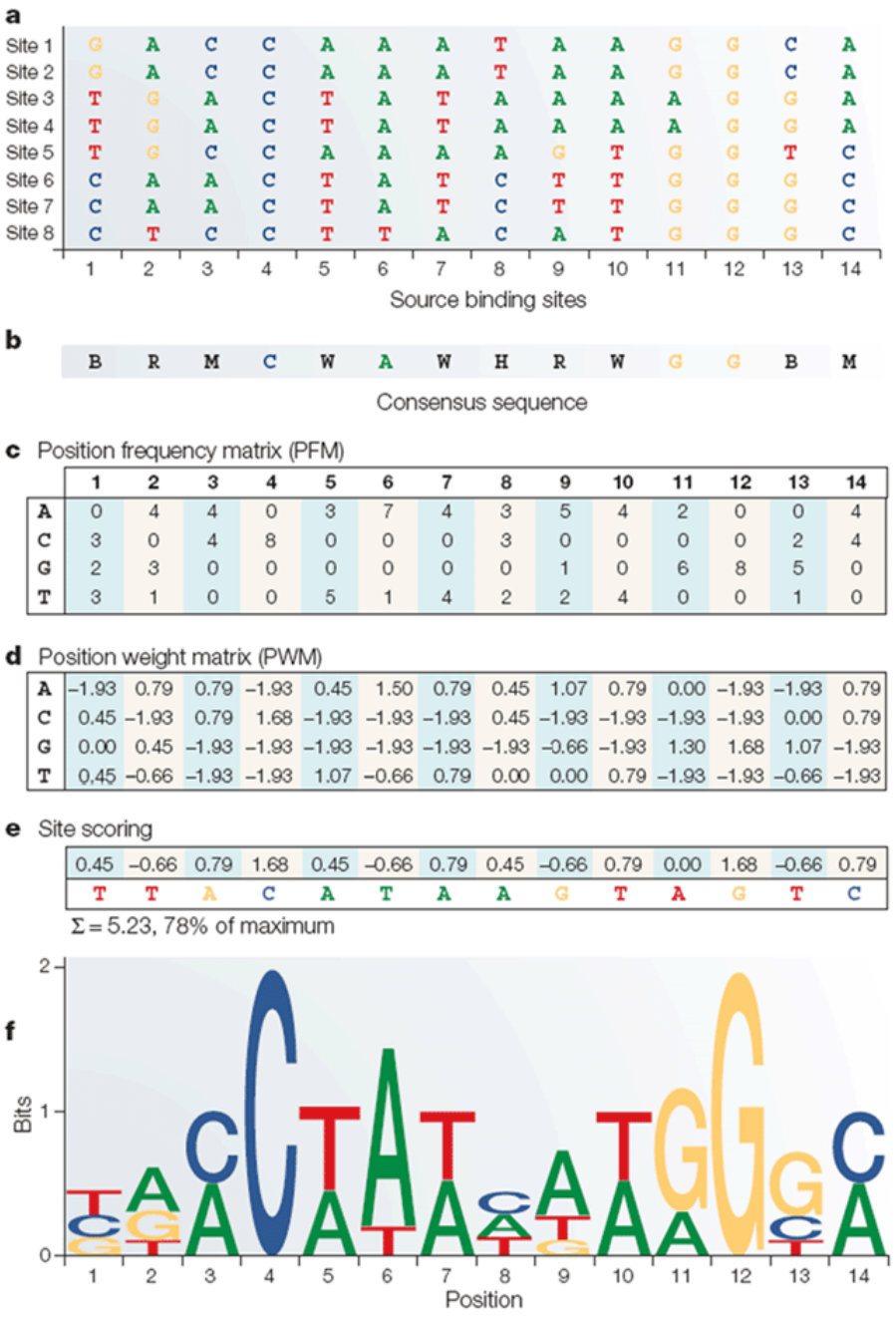


Figure 33. Models of TFBSs.
This figure is extracted from (Wasserman and Sandelin, 2004).

A set of experimentally validated 8 MEF2-binding sites is first collected from the literature and aligned (Figure 33 a). The sequence variability of the collection of binding sites strongly affects the subsequent models for predicting additional sites. The diversity between the sites is very important. For instance, 6 out of 14 nucleotides (~42%) are identical between sites 1 and 8.

Consensus sequences can be used to represent known TFBSs. A consensus nucleotide letter is assigned to represent the nucleotide composition in each column (Figure 33 b).Unusual binding sites can have an extreme effect on the consensus (see, for example, site 2).

To allow more flexibility, consensus sequences are often represented with the IUPAC nucleotide code (Table 9).

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	Purine: A or G
Y	Pyrimidine: C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Table 9. IUPAC nucleotide code.

A double-degenerate code (S, W, K and M) indicates that the corresponding two nucleotides occur in more than 75% of the aligned sequences but each of them is present in less than 50%. Usage of triple-degenerate codes (B, D, H and V) is restricted to those positions where one of the nucleotides did not appear at all in the sequence set and none of the afore-mentioned rules applies. All other frequency distributions are represented by the letter "N".

Although the use of consensus sequences provides better representation than a single sequence, it fails to reflect the quantitative characteristics of TFBSs.

To more accurately reflect the characteristics at each position, a Position Frequency Matrix (PFM) can be calculated. A PFM contains the number of observed nucleotides at each position of the motif (Figure 33 c). For instance, the first column in the alignment (Figure 33 a) consists of no A, three C, two G and three T, resulting in the corresponding first matrix column {0,3,2,3}.

A normalized PFM, in which each column adds up to a total of one, is a table of probabilities for observing each nucleotide at each position. The normalized frequency matrix is usually converted to a position weight matrix (PWM) by converting normalized frequency values to

a log-scale (Figure 33 d). PWMs are also known as position-specific scoring matrices (PSSMs, pronounced 'possums').

Using a matrix model, a quantitative score for any DNA sequence can be generated by summing the values that correspond to the observed nucleotide at each position (Figure 33 e). For large and representative collections of binding sites, the scores are proportional to binding energies (Stormo, 2000).

The specificity in each column of the alignment can be measured in terms of information content (Schneider and Stephens, 1990). A sequence logo scales each nucleotide by the total bits of information multiplied by the relative occurrence of the nucleotide at the position (Figure 33 f). Sequence logos enable fast and intuitive visual assessment of pattern characteristics.

4.3.2 TFBSs prediction algorithm

A gene may be regulated by a particular TF if its promoter contains the binding sites of this TF. Indeed, several programs have been developed to predict TFBSs on a genomic sequence on the basis of PWMs libraries such as MATRIX SEARCH (Chen et al., 1995), SIGNALSCAN (Prestridge, 1996) or TESS [see (Stoeckert et al., 1999)]. Nevertheless, MatInspector (Quandt et al., 1995) and Match™ (Kel et al., 2003) programs are the most used.

MatInspector uses similar scoring method as Match™ but uses different PWMs matrices is limited to an online access of a maximum of 20 requests free.

Match™ is a weight matrix-based tool for searching putative transcription factor binding sites in DNA sequences (Kel et al., 2003). Match™ is closely linked to and distributed together with the TRANSFAC® database. In particular, Match™ uses the matrix library collected in TRANSFAC®. Several sets of optimised matrix cut-off values are built in the system to provide a variety of search modes with different stringencies. Indeed, several profiles allow the user to minimise the number of false positives, false negatives or both. Other profiles defining matrices of high quality are also defined in the TRANSFAC® database and can be used as options in the Match™ program. Furthermore, a number of tissue-specific or taxa-specific profiles are also provided. The search algorithm of Match™ uses two score values: the matrix similarity score (MSS) and the core similarity score (CSS). These two scores measure the quality of a match between the sequence and the matrix, which ranges from 0.0 to 1.0, where 1.0 denotes an exact match. The core of each matrix is defined as the first five most conserved consecutive positions of a matrix.

4.3.3 Databases describing TFBSs

Several databases describing TFBSs have been developed. Currently, position weight matrices of many characterized TFs are available in databases such as JASPAR (Sandelin et al., 2004a) and TRANSFAC (Matys et al., 2006) for mammalian organisms.

4.3.3.1 JASPAR

JASPAR is a collection of transcription factor DNA-binding preferences, modelled as PWMs.

The current version of JASPAR contains three sub-databases (Vlieghe et al., 2006) presented in Table 10.

Data collection	JASPAR CORE	JASPAR FAM	JASPAR phyloFACTS
Keywords	Non-redundant, literature curated models (Sandelin et al., 2004a)	Meta-models for structural classes of TFs (Sandelin and Wasserman, 2004)	Data-mined profiles using phylogenetic pattern finding (Xie et al., 2005)
Number of models	123	11	174

Table 10. JASPAR sub-databases description.

The JASPAR CORE sub-database is an expanded version of the original, non-redundant collection of annotated, high-quality matrix-based transcription factor binding profiles. It contains a curated, non-redundant set of 123 profiles from published articles. All profiles are derived from experimentally defined TFBSs for multicellular eukaryotes (Sandelin et al., 2004a). As far as possible, the collection is non-redundant. JASPAR CORE should be used for the detection of putative binding sites resembling target sequences of known TFs.

The JASPAR FAM sub-database consists of models describing shared binding properties of 11 major structural classes of transcription factors. The collection facilitates prediction of TF binding domain structures based on profile information alone (Sandelin and Wasserman, 2004). Since many factors have similar target sequences, multiple predictions at the same locations that correspond to the same site are often observed. This type of model reduces the complexity of the results. The models are especially suitable for gene- and genome-wide exploratory searches in cases where there is no prior knowledge of cognate factors.

The JASPAR phyloFACTS sub-database contains 174 matrices computationally derived from statistically overrepresented, evolutionarily conserved sequences in the regulatory region of mammalian genes. The profiles were based on a comprehensive systematic survey of regulatory motifs (Xie et al., 2005), which used the phylogenetic relationship between

human, mouse, rat and dog to discover conserved and overrepresented sequence motifs in the region 2 kb upstream and downstream from the RefSeq-based TSS of human genes. These matrices are a mix of known and as yet undefined motifs. They are useful when other factors might determine promoter characteristics, such as structural aspects and tissue specificity. They serve as a non-redundant extension to JASPAR CORE.

The latest release of JASPAR is available at http://jaspar.genereg.net.

4.3.3.2 TRANSFAC

The database TRANSFAC® currently integrates several databases; we will focus on the Transfac database of eukaryotic *cis*-acting regulatory DNA elements and *trans*-acting factors (Matys et al., 2006). TRANSFAC® started in 1988 with a printed compilation (Wingender, 1988) and was converted into computer-readable format in 1990. The primary data of DNA-binding sites in TRANSFAC are based on experimental evidence. These data are extracted by curators from peer-reviewed articles. The curators search the scientific literature for suitable data, which are then entered via an input client, making use of controlled vocabulary and various automated functions, into a relational database, from which flatfile releases are generated from time to time. Collection of these data in a structured form allows the inference—via comparison and classification— of secondary or so-called meta-data (e.g. nucleotide distribution matrices, factor classification). Both types of data, the primary as well as the secondary data, can then serve as a basis for (sequence-based) predictions by certain programs, e.g. Match™ (Wingender et al., 1996) (for matrix-based transcription factor binding site searches). The Transfac data are stored in six tables which are described in Table 11.

Table	Description
FACTOR	describes proteins that regulate transcription (by sequence-specific interaction with DNA)
SITE	gives information on individual (regulatory) protein binding sites (within eukaryotic genes)
GENE	gives a short explanation of the gene where a site (or group of sites) belongs to
MATRIX	contains nucleotide distribution matrices for the binding sites of transcription factors
CELL	gives brief information about the cellular source of proteins that have been shown to interact with these sites
CLASS	contains background information about the transcription factor classes

Table 11. Transfac tables and their descrition.

The primary data are the three tables FACTOR, SITE and GENE for information on transcription factors, their binding sites and regulated genes, respectively. Nucleotide

distribution matrices, which are derived from a collection of binding sites for a particular factor, are stored in the MATRIX table.

In the public Transfac database release 7.0, the MATRIX table contains 398 nucleotide distribution matrices of aligned binding sequences. These sequences have been obtained by in vitro selection studies or may be compiled sites of genes. TRANSFAC PWM and matrix entry are detailed in Annexe 1 - .

The CLASS table groups the transcription factors according to their DNA-binding domains. In addition to this CLASS table, the factor entries are linked to the respective nodes in a classification hierarchy (Wingender, 1997). In a sixth table (CELL), cell lines and other kinds of factor sources, which were used for detection of a binding site/binding activity, are presented (Table 12).

Table	**TRANSFAC® Release 7.0**
FACTOR	6,133
Homo sapiens	1,040
Mus musculus	765
Drosophila melanogaster	233
Arabidopsis thaliana	1,751
Saccharomyces cerevisiae	368
SITE	7,915
MATRIX	398
Gene (all entries)	2,397
Homo sapiens	608
Mus musculus	417
Drosophila melanogaster	145
Arabidopsis thaliana	115
Saccharomyces cerevisiae	195
GENE (entries with SITE links)	1,504
CLASS	50
CELL	1,307

Table 12. Number of entries in the table of TRANSFAC® 7.0.

The Transfac database, release TRANSFAC 7.0 is a public version, accessible from http://www.gene-regulation.com/pub/databases.html . This Transfac database version is a less well maintained; data reduced version and is free-of-charge. The Transfac database also exists as a Professional version.

Statistics for the current professional version 10.1 are presented in the Table 13 and Table 14.

Table	Number of entries
Factor	8,021
Site	17,492
Gene	13,147
Matrix	795
Cell	2,131
Class	57

Table 13. General statistics for Transfac Professional release 10.1

Taxa	Number of entries
Vertabrata	569
Insecta	67
Plants	95
Fungi	56
Nematoda	7
Prokaryota	1

Table 14. Statistics on the matrix entry of Transfac Professional release 10.1.

While TRANSFAC is the most comprehensive of TF databases, it is still far from complete, containing binding sites for only a fraction of the known and putative transcription factors (Chang et al., 2006).

In principle, examination of genomic sequences with these PWMs should allow for the identification of TFBSs, and hence, regulatory regions. Nevertheless, as it has been described in this introduction, the size of the genomes, the promoter length which can extend to many kilobases from the TSS, combined with the fact that TFBSs are short and degenerated sequences which produce a large number of false-positive results, complicates this task enormously. Indeed, these motifs can be found everywhere for example in the human genome and experiments have shown that only an extremely small proportion represent *bona fide* TFBSs (Blanchette et al., 2006) (Chang et al., 2006). As a consequence, a possible solution to this problem is to identify a "sheltered environment" in which specificity of pattern discovery might be enhanced and therefore in which the number of false positives would be minimized. Thus, evolutionary information comes to the rescue (Corcoran et al., 2005) with two widely used methods: the phylogenetic shadowing and footprinting.

4.4 Phylogenetic shadowing

Boffelli *et al.* developed a method which was termed phylogenetic shadowing (Boffelli et al., 2003). Phylogenetic shadowing examines sequences of closely related species and takes into account the phylogenetic relationship of the set of species analyzed. This approach enabled the localization of regions of collective variation and complementary regions of conservation, facilitating the identification of coding as well as noncoding functional regions.

Thus, phylogenetic shadowing has emerged as a strategy for deciphering functional elements in comparisons of closely related species (such as different primates) (Boffelli et al., 2003). This method offers a variation on the basic principle: "What is important is conserved" and can be defined as "what is not critical can vary—at least some of the time". Standard pairwise comparisons between such sequences (that usually display 95% or higher sequence identity) fail to discriminate between slow- and fast-evolving regions due to a very low density of mutations. Phylogenetic shadowing overcomes this problem by comparing many closely related sequences simultaneously and combining mutations from all the sequences into a single conservation profile (Boffelli et al., 2003) (Ovcharenko et al., 2004). If the mutations occur independently in different lineages, they would be differently distributed in different sequences. Therefore, combining sequence mismatches from N different closely related sequences increases the divergence rate by a factor of N, thus allowing the separation of slow- and fast-mutating regions (Ovcharenko et al., 2004).

The work of Boffelli *et al.* (Boffelli et al., 2003), describes the phylogenetic shadowing of 17 primate species closely related to *Homo sapiens*, spanning 40 million years of evolution. Close examination of the sequence differences among these primate species revealed that although similarity is the rule, absolute conservation is the exception. Summing these exceptions reveals that the coding exons (as expected) as well as multiple regions smaller than typical exons (which may be regulatory elements) are highly conserved (Figure 34).

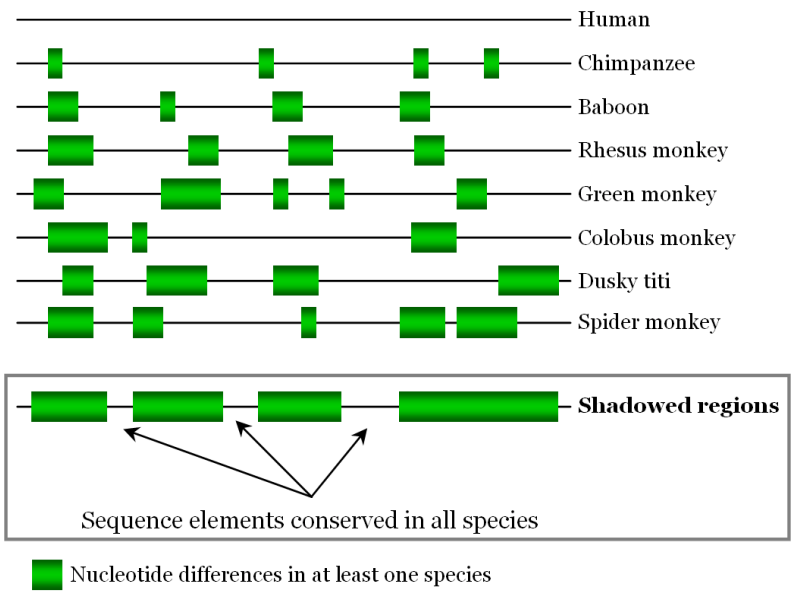


Figure 34. Primates in shadowland.
Phylogenetic shadowing enables multiple comparisons among DNA sequences from closely related primate species including human (Boffelli et al., 2003). In this way, the least variable regions of the genome, which should include exons and regulatory elements, can be identified. Figure extracted from (Gibbs and Nelson, 2003).

The authors experimentally analyzed several of these candidate regulatory regions with protein binding tests and gene reporter assays. They found that the predicted DNA regulatory sequences bound to nuclear proteins and enhanced transcription in reporter constructs. At least part of the reason for the success of phylogenetic shadowing is that the sum of the evolutionary distances spanned by several close relatives is as great as that between two distant species. This suggests that in-deepth sequence comparisons of numerous primate species should be sufficient to identify important regions of conservation that encode functional elements (Boffelli et al., 2003).

Despite the accuracy of this method, it is rarely used because of the lack of availability of completely sequenced genomes of closely related species such as sequences of primates.

4.5 Phylogenetic footprinting

4.5.1 Description of the method

Comparative genomics is a powerful approach for the extraction of biologically meaningful information, the prediction of gene structure and of *cis*-acting regulatory elements (Cooper and Sidow, 2003) (Frazer et al., 2003) (Ureta-Vidal et al., 2003). Indeed, 'phylogenetic footprinting' is a widely applied approach to identify regulatory regions and potential

transcription factor binding sites (TFBSs) using alignments of non-coding orthologous regions from two or more organisms (Sauer et al., 2006). This method was first introduced by (Tagle et al., 1988) who investigated primate γ- and ε-globin genes.

The basic assumption of phylogenetic footprinting is that regulatory elements in non-coding regions are under a higher selective pressure during evolution than non-functional regions. Deleterious mutations in regulatory regions are removed from a population owing to selection and non-functional regions are therefore expected to accumulate more mutations over time than regulatory regions (Sauer et al., 2006). Aligning orthologous regions from two or more organisms hence should highlight the conserved parts which are expected to play a functional role (Figure 35). The term 'functional' is used in the following to mean an experimentally proven TF–DNA interaction.

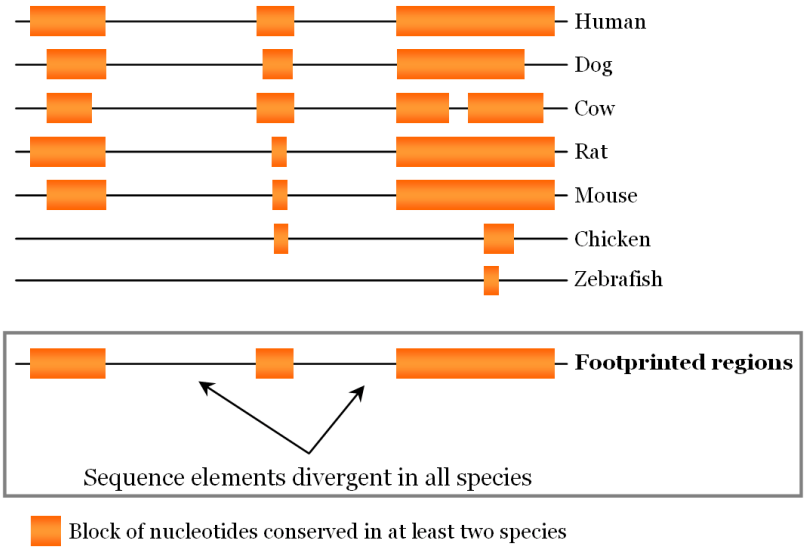


Figure 35. Phylogenetic footprinting.

Thanks to the availability of an increasing number of complete eukaryotic genomes, phylogenetic footprinting is now widely applied to identify regulatory regions and potential transcription factor binding sites (TFBSs) in different organisms. The inherent problem of comparative genomics is the choice of species to be compared with each other in order to reliably identify functional regions in the genomic sequence.

Interest in phylogenetic footprinting was increased when comparison of the mouse and human genomes showed that a surprisingly high proportion of the genome could be aligned and that at least 1.5% of the genome was highly conserved non-repeat, non-protein coding DNA (Waterston et al., 2002). Thus, human–mouse comparisons have been extensively used to identify potential regulatory regions which in many cases proved to be functional (Elnitski et al., 2003). Human and mouse diverged roughly 75–90 million years ago from their last

common ancestor. The divergence rate between the human and mouse genomes is low enough that one can still align orthologous sequences, but has provided sufficient time for a large fraction of nucleotides to have been exposed to considerable mutation and selection pressure (Boffelli et al., 2003), allowing the discrimination of functional elements based on their greater conservation. To assess the usefulness of human–mouse comparisons, several studies have been performed to determine to what extent experimentally known TFBSs can be identified by phylogenetic footprinting (Lenhard et al., 2003) (Liu et al., 2004). The data collections of these studies included betweeen 99 and 481 TFBSs of which about 60–68% could be detected by human–mouse comparisons.

As the DNA-binding specificity of some TFs is low, sequence conservation may not reflect properly the existence of an orthologous TFBS. Furthermore, several studies found that between 60 and 72 % of experimentally defined TFBSs are present in regions conserved between human and mouse (Sauer et al., 2006). The boundaries between exonic and intronic sequences can be considered as transition points from the regions where most of the sequences play biologically significant roles to the regions where most of the sequences are biologically less relevant. Similarly, it has been suggested that, in the promoters, most of the biologically significant elements should be embedded inside rather than outside the conserved blocks described in section 3.6.2 (Suzuki et al., 2004a).

At the time of writing, several projects of mammalian whole-genome sequencing are ongoing, extensive phylogenetic comparative analyses using the genomic sequences of these mammals should lead to more accurate phylogenetic footprinting results (Suzuki et al., 2004a).

As seen in section 1.3, TFBSs are weak patterns that become impossible to distinguish within long candidate regulatory regions. Cross-species comparison of sequences from orthologous genes, or phylogenetic footprinting reduces the amount of sequence under consideration by focusing attention on conserved regions that are more likely to serve a biological function (Boffelli et al., 2003). Although such methods can increase binding-site densities by fivefold, only the strongest sites are detected at this level (Thompson et al., 2004b).

The basic assumption of phylogenetic footprinting is that most functional regulatory elements (TFBSs) are conserved during evolution. Although functional elements may exist in nonconserved sequence, they are most likely to be found in regions conserved across species (Chang et al., 2006). The fraction of TFBSs present in nonconserved sequence is species-specific and will not therefore be detectable using such sequence comparison analysis approaches.

There are several crucial factors in phylogenetic footprinting which have to be addressed in order to obtain reliable results. The three major problems are (1) the identification of the correct orthologous sequence from the second organism, (2) the choice of an alignment algorithm and (3) the choice of criteria for determining conservation.

4.5.2 Orthologous sequence identification and localization on the genome

Retrieving the correct orthologous sequences is a crucial factor in *in silico* promoter analyses that take into account evolutionary information. This step requires previous knowledge and, orthology, paralogy have to be taken into account as well as current knowledge concerning the relative distance to the TSS in orthologous genes.

4.5.2.1 Orthologs/Paralogs

- Two genes are defined to be homologous if they have a common ancestor. Homology of sequences can be of two types: orthology or paralogy. Orthologs are defined as any pair of genes whose ancestor node is a speciation event: if a gene exists in a species, and that species diverges into two species, then the copies of this gene in the resulting species are orthologous. Paralogs correspond to any pair of genes whose ancestor node is a duplication event: if a gene in an organism is duplicated, then the two copies are paralogous. Furthermore, several situations of orthology/paralogy can be observed: orthology "one2one", "one2many" or "many2many" and paralogy either within or between species. These notions of different orthology and paralogy are illustrated in Figure 36.

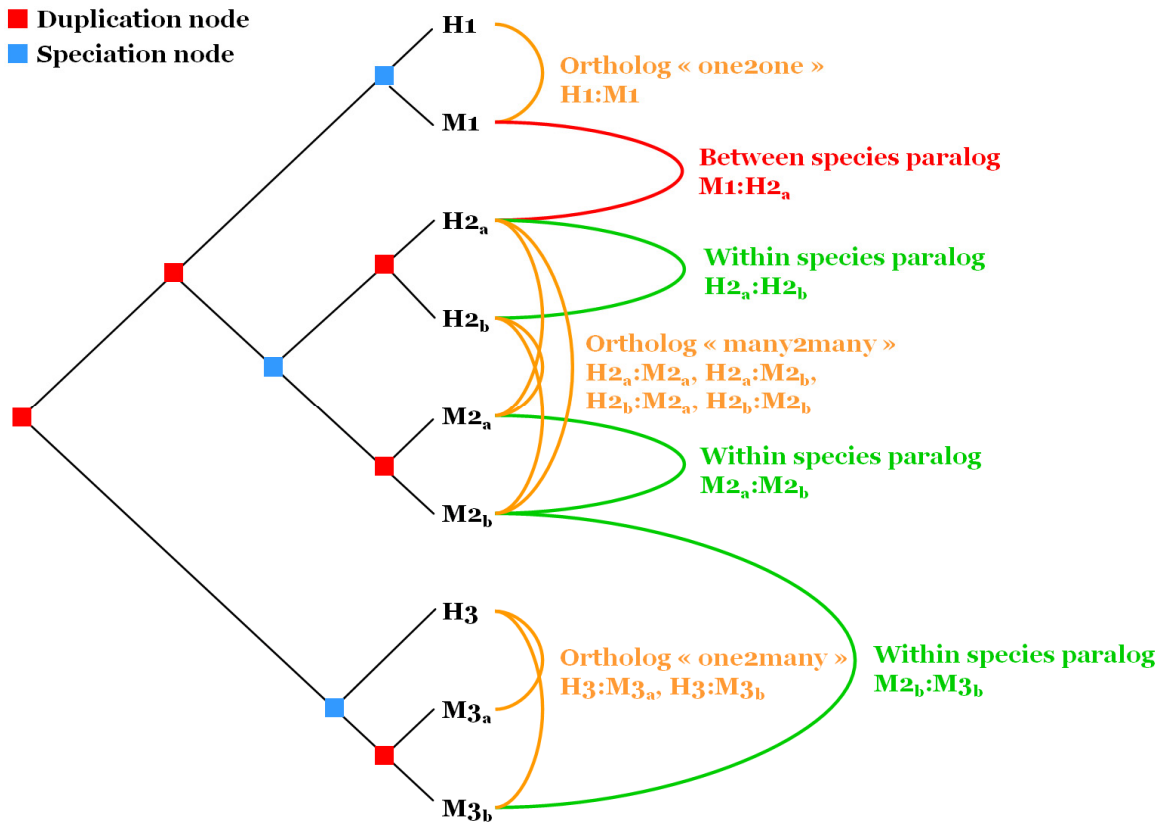


Figure 36. Sequence homology: orthology and paralogy
Schematic gene tree containing 2 species H and M (numbers point to different genes). This figure has been extracted from the Ensembl web server http://www.ensembl.org/info/data/compara/homology_method.html

"Within species paralogs" correspond to 2 genes of the same species whose ancestor node has been labelled as a duplication node e.g. H2a:H2b, M2a:M2b but does not necessarily mean that the duplication event has occurred in this species only. For example, M2b:M3b are also "within species paralogs" but the duplication event has occurred in the common ancestor between species H and species M. If H is human and M Mouse, the taxonomy level "times" the duplication event to the ancestor of "Euarchontoglires". "Between species paralogs" corresponds to 2 genes of different species whose ancestor node has been labelled as a duplication node e.g. M1:H2a or M1:H3.

- Orthologs will typically have the same or similar function. This is not always true for paralogs since due to lack of the original selective pressure upon one copy of the duplicated gene, this copy is free to mutate and acquire new functions.

At least 80% of mouse genes have only a single identifiable homologous gene in the human genome, which should be an ortholog (Waterston et al., 2002).

4.5.2.2 Pseudogenes

Pseudogenes are DNA sequences that are relatives of known genes but most of the time they are not transcribed due to the absence of a functional promoter. The human genome is the genome containing most pseudogenes.

Pseudogenes are complete or partial gene copies. They have lost, from their initial definition, their protein-coding ability (Mighell et al., 2000) (Vanin, 1985). There are two classes of pseudogenes: nonprocessed (duplicated) and processed Figure 37.

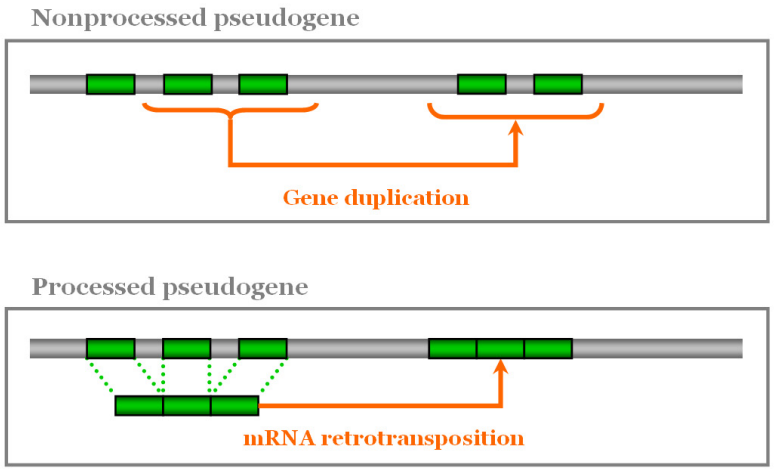


Figure 37. Nonprocessed and processed pseudogenes

Nonprocessed pseudogenes arose from complete or partial genes, so they have retained the identical, though often incomplete, exon structure of the original functional gene. Only a small fraction of these pseudogenes keep their genic functionality duplication (Prince and Pickett, 2002).

Processed pseudogenes result from mature mRNA retrotransposition, i.e. reverse transcription of an mRNA transcript followed by random integration into genomic DNA. Thus, they do not contain a promoter region. Nevertheless, some « active » genes, intron less and originating from a retrotransposition have been described in many organisms (Brosius, 1999). Such pseudogenes are so-called retrogenes (Torrents et al., 2003).

In the human genome, 9747 processed and 9546 nonprocessed pseudogenes have been described (Zhang et al., 2004). About 40% of the processed pseudogenes are located in syntenic regions conserved between human and mouse. This suggests that they were created before the divergence of human and mouse.

4.5.2.3 Relative distance to the TSS in orthologous genes

Intuitively orthologous sequences from a second species should be retrieved according to the annotated TSS, e.g. Jareborg *et al.* (1999) (Jareborg et al., 1999) or Liu *et al.* (2004) (Liu et al., 2004) focused their studies on the sequences 1000 bp upstream of the TSS for both human and mouse. Nevertheless, retrieving an orthologous sequence from the second species using only the identical relative distance to the TSS in the first genome may not identify the correct orthologous region, since the TSS in the second species may be different or may be annotated incorrectly in one of the genomes. Indeed, Frith *et al.*, have shown that the TSS of orthologous genes do not always reside at equivalent positions in the human and mouse genomes (Frith et al., 2006). They also showed that while most TSSs in mouse have a counterpart at the homologous location in the human genome sequence and *vice versa*, a significant number of cases of "TSS turnover" can be observed, in which corresponding TSSs in human and mouse have been translocated by more than100 bp. These range from extreme cases in which a TSS is observed in one species but not in the other to intermediate instances in which a gene has two alternative start sites that are differentially dominant in each species (Frith et al., 2006). The existence of "TSS turnover" has implications for phylogenetic footprinting as well as promoter prediction that rely on cross-species comparison analysis.

Moreover, when analyzing enhancer sequences that may be located at relatively large distances from the TSS, the chances of insertion/deletion between enhancers and the TSS increase with the distance (Sauer et al., 2006).

4.5.3 Alignment of noncoding genomic sequences

Another crucial step in promoter sequence comparison during evolution is the alignment of these noncoding genomic sequences. Several multiple alignment programs exist but they are not all specifically designed for aligning genomic and noncoding genomic sequences. Although some computer programs designed primarily for aligning multiple protein sequences (Thompson et al., 1994) can be applied to DNA sequence, the premise that DNA presents distinct challenges and opportunities has motivated the development of several multiple alignment programs specifically for genomic sequences (Blanchette et al., 2004).

Some tools for the alignment of noncoding DNA are presented in this section. They have been selected upon a number of criteria such as their public availability and the possibility to download the program for local installation.

The general properties and characteristics of these programs are presented in Table 15.

Program	Alignment		Program characteristics
	Pairwise/ Multiple	**Global/Local**	
ClustalW	Multiple	Global	Optimized for aligning multiple protein sequences Progressive alignment program Too slow for alignment of sequences of tens kb
DIALIGN		Local and Global outputs	Quite good for divergent sequences Too slow for alignment of sequences of tens kb
CHAOS	Pairwise	Local	Chains together pairs of similar regions, one from each of the two input DNA sequences
LAGAN	Pairwise	Global	Based on CHAOS program Rapid global alignment of two homologous genomic sequences
MLAGAN	Multiple	Global	Based on progressive alignment with LAGAN program
AVID	Pairwise	Global	Relies on "anchoring" approach
MAVID	Multiple	Global	Based on progressive alignment with AVID program Aligns multiple genomic regions up to megabases long
BLASTZ	Pairwise	Local	Independent implementation of the Gapped BLAST
MULTIZ	Multiple	Local	Creates "reference sequence" alignment Can be used with sequences that are fragmented or have rearrangements such as inversions or duplications Used to build whole-genome alignments for the UCSC Genome Browser
TBA	Multiple	Local (Threaded blockset)	Based on BlastZ and MULTIZ programs Finds matching regions that occur in the same order and orientation in all species Does not accommodate inversions or duplications Consider a phylogenetic tree as input Good for divergent sequences

Table 15. General characteristics of programs used for noncoding DNA alignment.
A global alignment entirely covers the given sequences whereas a local alignment covers only part of the sequences. A pairwise alignment is an alignment between two sequences while a multiple alignment involves more than two sequences.

The general approaches of the alignment algorithms used in each of these programs are briefly described in Annexe 2 - .

Whereas all these methods are generally successful in aligning human and primate sequences, TBA and DIALIGN give better results than the others on more divergent sequences such as human and rodent. MULTIZ leads to high quality alignments but produces reference sequence alignments.

Many of the programs presented above have been evaluated in terms of sensitivity and specificity (Blanchette et al., 2004). Sensitivity and specificity are particularly useful to determine whether a given program over- or under- predicts aligned bases. The sensitivity is defined as the fraction of aligned bases of the correct alignment that are paired identically in the predicted alignment, while the specificity is defined as the fraction of aligned bases of the predicted alignment that are paired identically in the correct one. Blanchette found that TBA,

MULTIZ, MLAGAN and DIALIGN had similar sensitivity and specificity, although TBA and MULTIZ obtained sensitivity and specificity slightly higher than the two other programs. The MAVID program has a better specificity than sensitivity, indicating that it tends to underpredict aligned bases, whereas CLUSTALW tends to overpredict aligned bases. Furthermore, in terms of running times, programs initially designed for aligning large genomic regions (MULTIZ, TBA, MAVID and MLAGAN) are the only ones to run fast enough to contemplate whole-genome alignments. Among these programs, MAVID stands out with a remarkably fast running time.

The increase in vertebrate genome sequencing will soon allow a systematic noncoding sequence analysis with a number of species much larger than the classical human/mouse pair. Such alignments are important because they should improve and refine the accuracy of the alignment and thus of the information in terms of conservation. This implies that there is still some room for improvement of the programs dedicated to noncoding sequence multiple alignments.

4.5.4 Description of tools dedicated to phylogenetic footprinting

Some tools dedicated to phylogenetic footprinting are presented in this section. They have been selected because they are the most used for promoter analysis. General properties of these tools are depicted in Table 16.

Program	Input sequences	Alignment programs	TFBS search
rVISTA (2002)	Two orthologous sequences	AVID	Match program on the TRANSFAC database
CONSITE (2003)	Pair of orthologous genomic sequences	ORCA aligner (Arenillas and Wasserman, unpublished)	JASPAR database
CONREAL (2004)	Two orthologous sequences or a gene name	CONREAL alignment program based on the use of TFBSs predictions as anchors Or LAGAN, BLASTZ and MAVID	JASPAR and TRANSFAC databases
rVISTA 2.0 (2004)	Two genomic sequences or precomputed alignments from GALA database or ECR Browser	zPicture and BLASTZ	tfSearch tool (I. Ovcharenko, unpublished data) TRANSFAC or user-defined consensus sequences
FOOTER (2005)	Single protein sequence or two homologous promoter sequences	DBA (Jareborg et al., 1999)	PSSM constructed from TRANSFAC binding sites

Table 16. Overview of the main phylogenetic footprinting programs.

The method used by each of these tools is further briefly described in Annexe 1 - .

4.6 Analysis of a group of co-expressed genes analysis

The study of the promoters for a group of co-expressed genes is another method allowing an increase of the signal-to-noise ratio.

Genome-wide mRNA-profiling experiments allow the identification of genes that have similar expression patterns. As co-expressed genes are likely to be regulated by the same TFs, it is thought that the analysis of noncoding sequences of coexpressed genes will be useful in identifying common cis-regulatory elements. These methods have been successfully applied to simple organisms such as yeast and worm (Hughes et al., 2000) (GuhaThakurta et al., 2002). They have also been applied to mammals. Indeed, calculation of the enrichment of a TFBS in a set of coregulated genes against a "reference" set such as randomly selected genes in the genome has been used to highlight common TFs regulating genes involved in the same biological process (Aerts et al., 2003) (Elkon et al., 2003). Some of these studies have nevertheless been unsuccessful in mammals because intergenic sequences in higher eukaryotes are very long and contain a large amount of nonregulatory sequences (Chang et al., 2006).

Furthermore, it is important to note that co-expression alone does not imply co-regulation. Genes may be expressed within a given type of cell through multiple and cascading responses to a single stimulus. Indeed, sets of genes that are likely to be co-regulated are, for example, subsets from large-scale expression studies that follow a consistent and common time course or that are involved in the same biological process (commonly defined with terms of Gene Ontology, see section7.1) (Thompson et al., 2004b).

Material and methods

Chapter 5 - Informatic and bioinformatics resources

As described in the introduction, the post-genomic era is generating a huge amount of data. In this context, bioinformatics sequence analyses use this large amount of data that is stored in a number of distinct databases and in addition the analyses also generate a large amount of data. As a result, current developments in bioinformatics require important computational infrastructure. The developments and analyses performed during this thesis used the existing infrastructure and computer resources of the BioInformatics Plateform of Strasbourg (http://bips.u-strasbg.fr/) of the IGBMC institute, where important informatics equipment with high storage and calculation capacity is available.

5.1 Informatics resources

5.1.1 Calculation and data storage possibilities

Three central servers are currently available for program development and computational data analyses:

- Interactive and web services: "Titus", a SUN quadriprocessor (Enterprise 450) with 1 Go shared memory.
- Computational servers: "Beaufort", composed of six Compaq ES40 cluster (Tru64 Unix) quadriprocessor, five have each 4 Go memory and the sixth has 16 Go. The second computational server, called "Star", is composed of 6 Sun Enterprise V40z servers. Two use the Solaris 10 exploitation system andare dedicated to storage and disk access. The other four use RedHat Enterprise Linux 4 and are dedicated to calculation. The total system consists of 6 x 4 opteron processors with 2 x 32 Go and 4 x 16 Go memory.
- Disk server: Sun V480 (Solaris 9) providing 8 Tera-bytes on Raid5 disks shared with other servers using NFS.

5.1.2 Programming languages

PromAn is developed in Tcl/Tk. This programming language has numerous advantages:

- Tcl/Tk is a script language, and so needs no compilation.
- This implies that it can be ported to any informatics system (PC, MAC, Unix, Linux).
- It is particularly adapted to file reading and handling as each variable in Tcl/Tk is a string.
- Tk is a programming language allowing the creation of complete graphical interfaces.
- Finally, Tcl/Tk is easy to use, versatile and widely used in the laboratory.

The PromAn web server also integrates html programming and Java scripts.

This research work has necessited bioinformatics resources such as databases and programs dedicated to genomic sequences and to protein sequences. The UCSC genome browser web-server and the NCBI HomoloGene database have been mainly used for promoter and orthologous promoter sequence retrieval. Local tools available in the laboratory have also been used, such as the GCG analysis platform and the PipeAlign suite of tools.

5.2 External web servers

5.2.1 Genome browser web servers

Several genome browsers are available via web servers such as NCBI (National Center for Biotechnology Information) (http://www.ncbi.nlm.nih.gov/Genomes/), Ensembl (http://www.ensembl.org/index.html), UCSC (University of California, Santa Cruz) (http://genome.ucsc.edu/). All these browsers provide similar facilities such as sequence alignment on genomes; genomic sequence retrieval; gene, transcript and protein mapping on genome etc. Nevertheless, UCSC provides numerous standardized flat files for all the species allowing an automatic data extraction. In addition, it is the only browser that provides a BLAT (Kent, 2002) server dedicated to map input sequences on genomes. BLAT, is more accurate and 500 times faster than popular existing tools for mRNA/DNA alignments. Moreover, UCSC provides a very fast genome sequence retrieval module that can be used with simple chromosome positions as input (chromosome, start, end and sense). In this section, this web server will be described.

The UCSC genome browser interface with the human rhodopsin gene is presented in Figure 38.

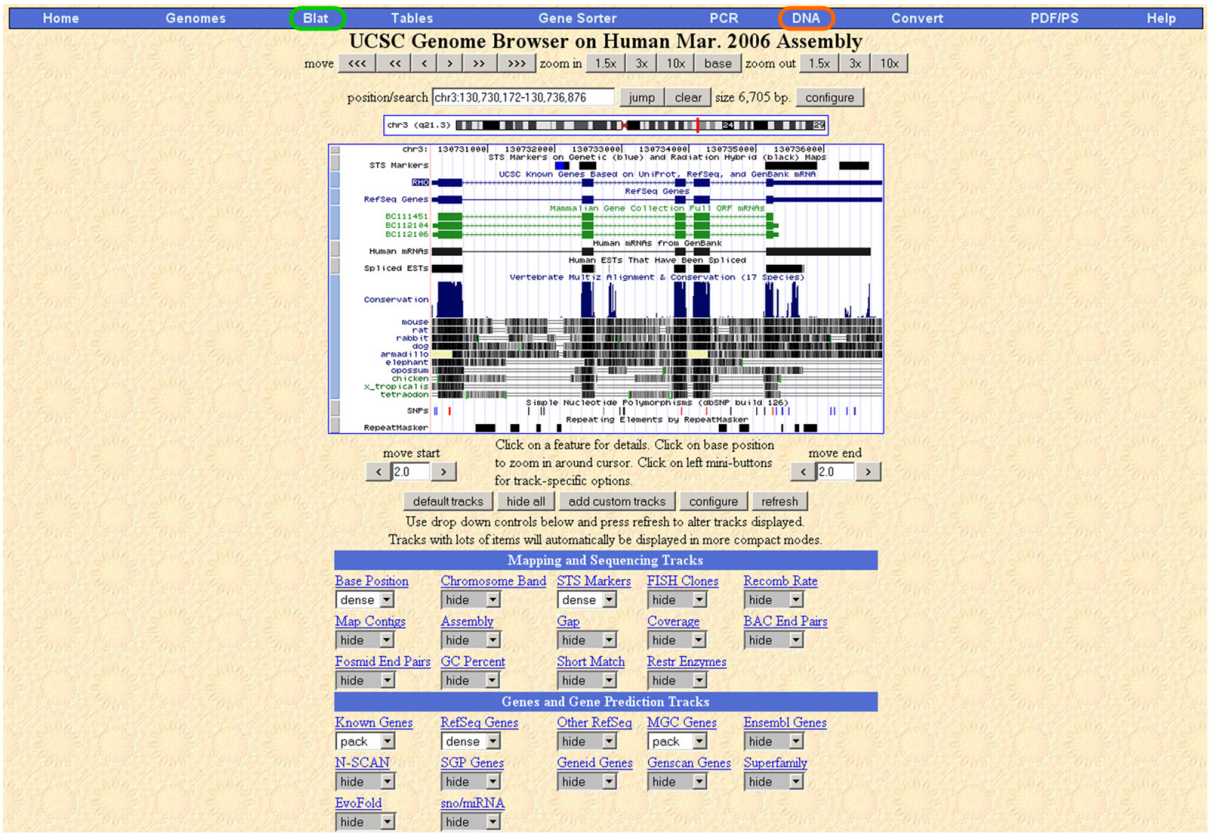


Figure 38. UCSC genome browser interface.
The human rhodopsin gene is shown in the UCSC Genome Browser. The mRNA, EST, the repeat localizatiuon as well as the conservation profile during evolution (from the MULTIZ alignment between 17 vertebrates) are depicted.

The UCSC Genome Browser provides a rapid and reliable display of any requested portion of genomes at any scale, together with dozens of aligned annotation tracks (known genes, predicted genes, ESTs, mRNAs, CpG islands, assembly gaps and coverage, chromosomal bands, mouse homologies etc).

The "DNA" link highlighted in orange in the previous figure gives access to a window allowing DNA extraction of the sequence previously visualized with the possibility of adding extra bases upstream and downstream (Figure 39).

Figure 39. UCSC, DNA sequence extraction.

Extraction of genomic sequences including the promoter sequence.

In addition, a given (amino or nucleic acid) sequence selected by the user can also be localized with the "Blat" alignement program (bordered in green in Figure 40).

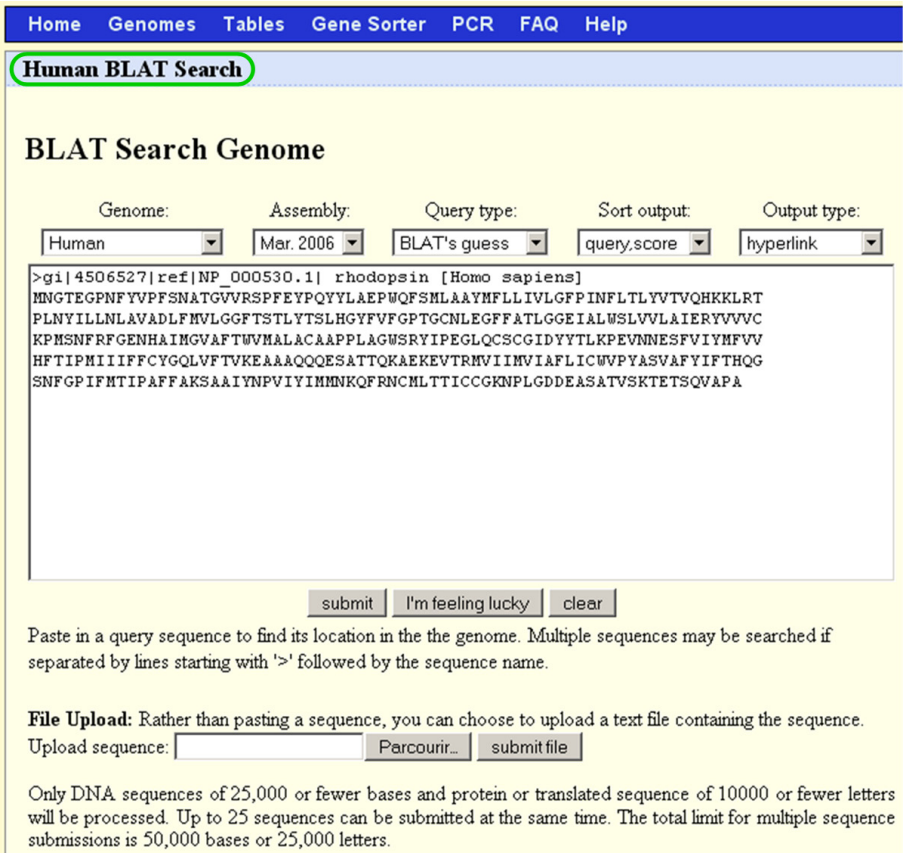


Figure 40. Blat program search

Blat (Kent, 2002) is a local alignment program dedicated to locating small sequences in large genomic sequences such as the location of cDNA in whole genomes. Afterwards, in the "Blat

Search Result" page (Figure 41), the link "browser" (highlighted in green) of the true genomic localization gives access to the UCSC genome browser illustrating the Blat query sequence given by the user.

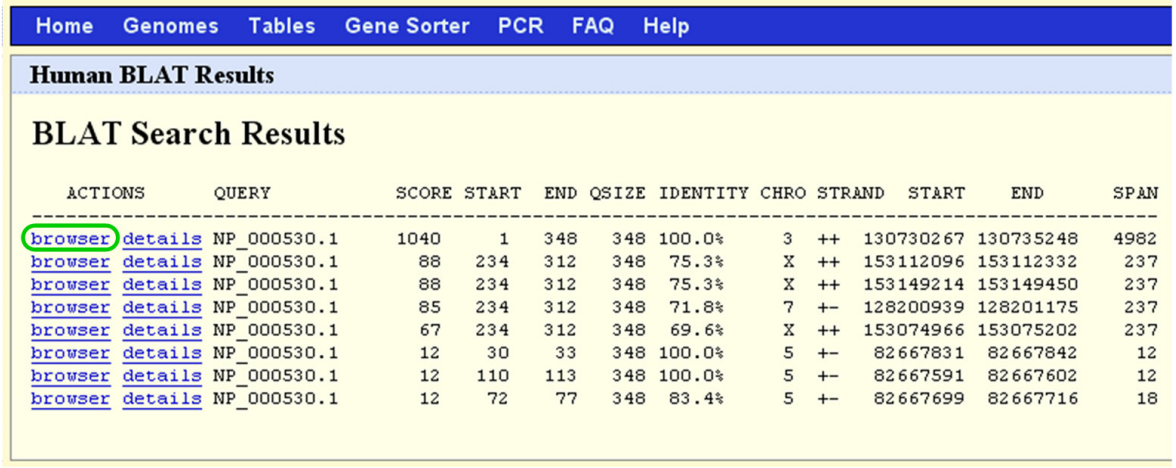


Figure 41. Blat search results
The first hit (outlined in green) was chosen as corresponding to the location of the rhodopsin protein on the human genome with a perfect identity (IDENTITY = 100.0%). Indeed, this hit covers the totality of the protein sequence (START = 1, END = 348 and QSIZE = 348). Furthermore, the coding exons are separated by introns on the genomic sequence because the SPAN (4982) is higher than the query size. This can also be confirmed by looking at the alignment with the "details" link.

Afterwards, the user can also extract a genomic sequence in the same way as explained above (Figure 38 and Figure 39).

5.2.2 NCBI HomoloGene database

The HomoloGene database (Wheeler et al., 2006) describes groups of homologous sequences that are automatically detected among the annotated genes of several completely sequenced eukaryotic genomes. These HomoloGene groups describe orthologous and paralogous sequences of a given gene. Statistics concerning the current version, HomoloGene release 51.1, are presented in Table 17.

Species	Number of genes		HomoloGene groups
	Input	Grouped	
H. sapiens	22,873	19,242	18,465
P. troglodytes	21,526	12,893	12,832
C. familiaris	19,766	16,687	16,252
M. musculus	24,175	20,502	19,136
R. norvegicus	21,991	19,006	17,780
G. gallus	18,029	12,233	11,405
D.melanogaster	14,017	8,105	7,900
A. gambiae	13,909	8,439	7,883
C. elegans	20,063	5,166	4,929
S. pombe	5,043	3,210	3,174
S. cerevisiae	5,863	4,738	4,588
K. lactis	5,335	4,453	4,422
E. gossypii	4,726	3,943	3,934
M. grisea	11,109	6,297	5,881
N. crassa	10,079	5,911	5,905
A. thaliana	26,659	11,168	10,846
O. sativa	33,553	11,053	9,460
P. falciparum	5,222	978	957
			165,749

Table 17. HomoloGene release 51.1 statistics
The initial number of genes from complete genomes, number of genes placed in a homology group, and the number of groups are described for each species.

The HomoloGene database is available at the following address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene .An example of the HomoloGene group number 68068, corresponding to the rhodopsin gene is presented in Figure 42.

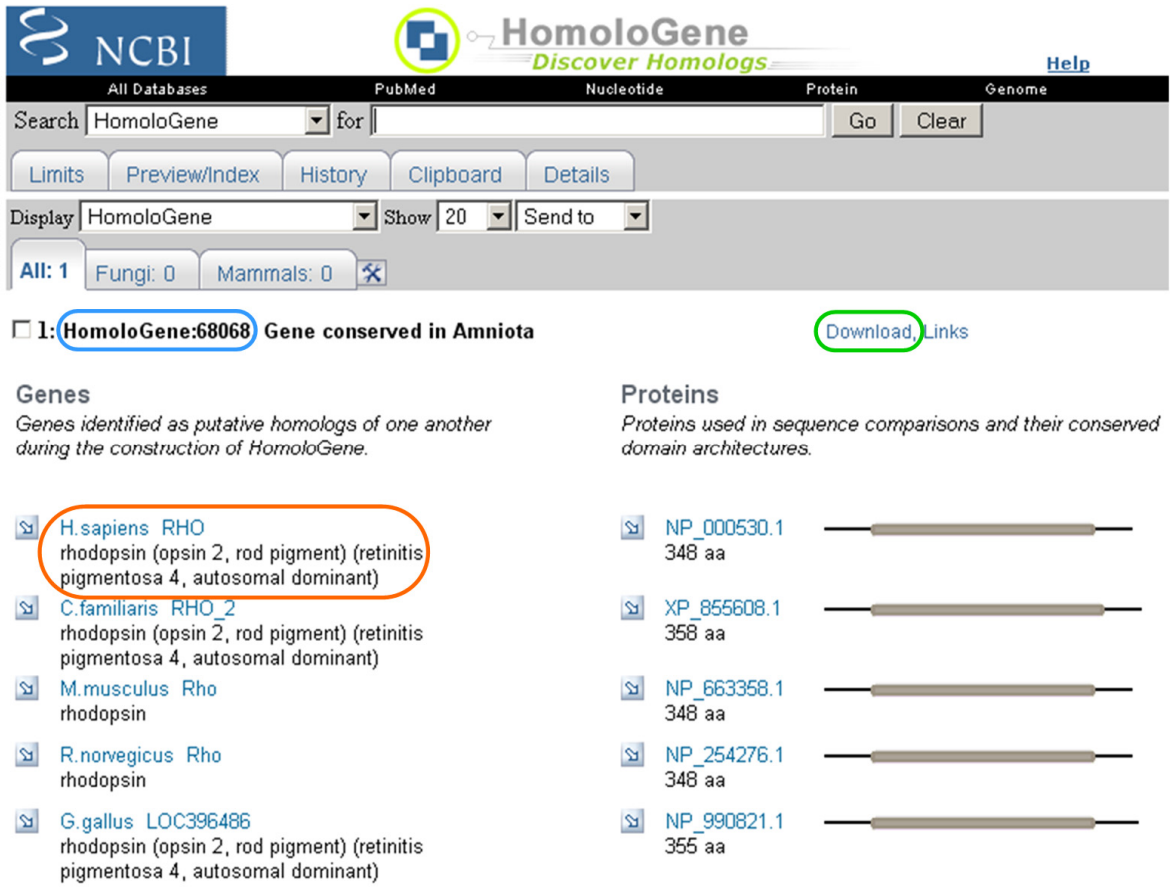


Figure 42. HomoloGene group ID number 68068, rhodopsin gene.

In this group, five orthologous sequences are described. They have been identified in the human, dog, mouse, rat and chicken genomic sequences. The link depicted in orange gives access to the Entrez Gene data related to the human rhodopsin gene (Figure 43).

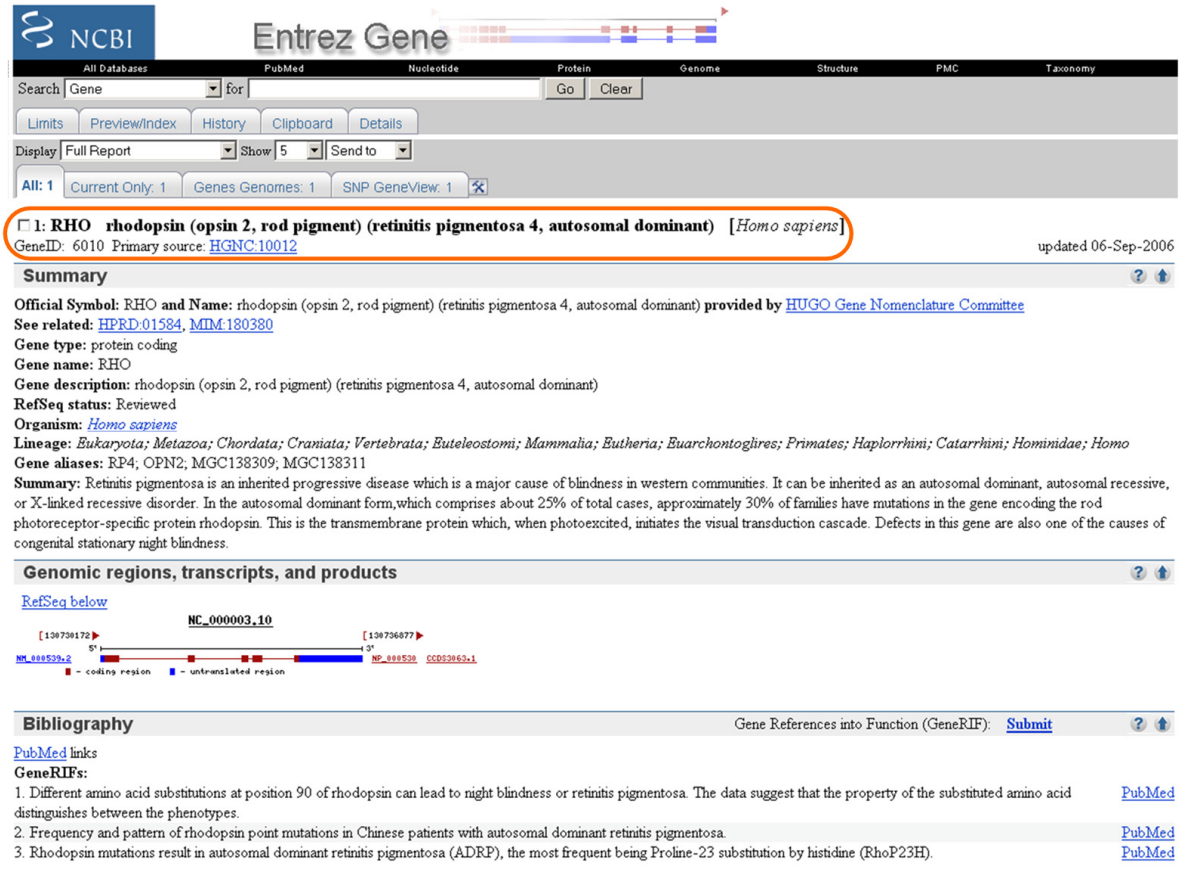


Figure 43. Link from HomoloGene database to Entrez Gene.

The NCBI also provides a tool to extract genomic sequences from the HomoloGene database ("Download" link bordered in green in Figure 42). The HomoloGene Downloader tool (Figure 44) is very useful as it allows the user to extract a group of orthologous genomic sequences in a single step.

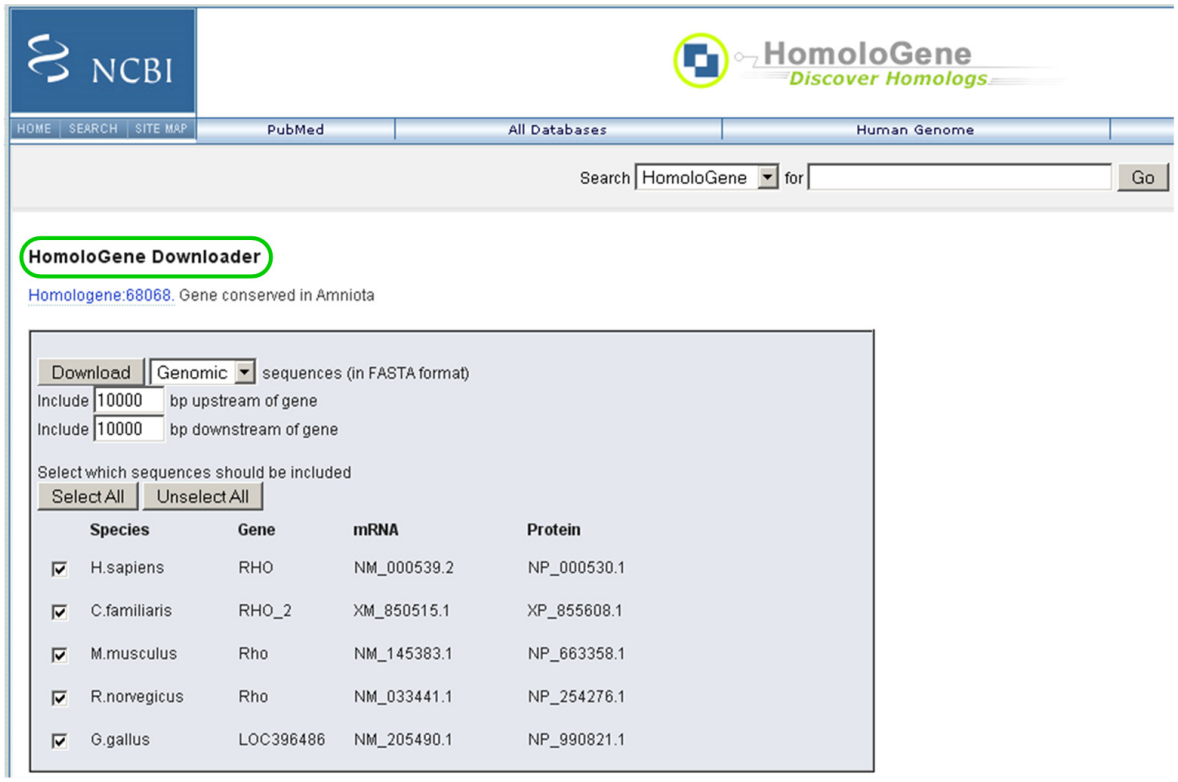


Figure 44. HomoloGene Downloader.
In the present example, genomic sequences extending 10 kb upstream and 10 kb downstream of the rhodopsin gene for the human, dog, mouse, rat and chicken are requested in fasta format.

The user has the possibility to define the nature (protein, mRNA or genomic sequence), the size, the extension upstream and downstream of the gene in the case of genomic sequence extraction and the species.

5.3 Local databases

A number of general as well as some more specialist databases are installed and updated regularly on the IGBMC servers. The databases used for this work are described in more detail below. The local databases are available in GCG format (Butler, 1998) and can also be queried using the SRS (Sequence Retrieval Software) system (Etzold and Argos, 1993).

5.3.1 Generalist databases

The main public sequence and structure databases have been installed locally on the IGBMC servers. The protein sequence database UniProt (Wu et al., 2006), consists of both SwissProt and SpTrEMBL databases (Boeckmann et al., 2003). The SpTrEMBL sequences are produced by automatic translation of the coding sequences from the EMBL nucleotide sequence

database. After validation and annotation by experts, the sequences in SpTrEMBL are incorporated in the SwissProt database.

5.3.2 Specialist databases

Our developments used the Gene Ontology database (Ashburner et al., 2000) (http://www.geneontology.org) and promoter dedicated databases (see Chapter 6 -). We also used the TRANSFAC professional database available in the laboratory (see 4.3.3.2).

5.3.3 Database query systems

5.3.3.1 GCG package

The GCG package (Wisconsin Package Version 10.2, Genetics Computer Group Madison, Wisc.) is a software suite containing diverse sequence analysis programs. This package allows the manipulation, visualization, analysis and comparison of sequences in the GCG format databases installed locally. In particular, several tools allowing sequence and multiple alignment format conversions have been used. They include the "fromfasta", "tofasta" tools allowing the conversion of a sequence in FASTA (TFA) format into GCG format and vice versa.

5.3.3.2 Sequence Retrieval Software (SRS)

Most of the local databases can be queried using the Sequence Retrieval Software (SRS) interrogation system. SRS currently allows access to more than 150 biological databases, including nucleic acid and protein sequences and structures, protein domains and metabolic pathways. It is designed for the extraction of semi-structured data, i.e. textual data with a pre-defined structure that may include redundancies or irregularities. The textual data is stored in flat files containing all the entries of a database. The flat files are organised into structured fields, which may be different depending on the databases. SRS performs a grammatical parsing of the information contained in the flat files and then indexes the different fields associated with each entry. This indexing allows a rapid access to the entry fields via complex queries, as well as cross-referenced queries that exploit the links between the different databases. Version 7 of SRS is currently installed at the laboratory. Database queries can be performed using UNIX commands or interactively (http://bips.u-strasbg.fr/srs/ , see Figure 45).

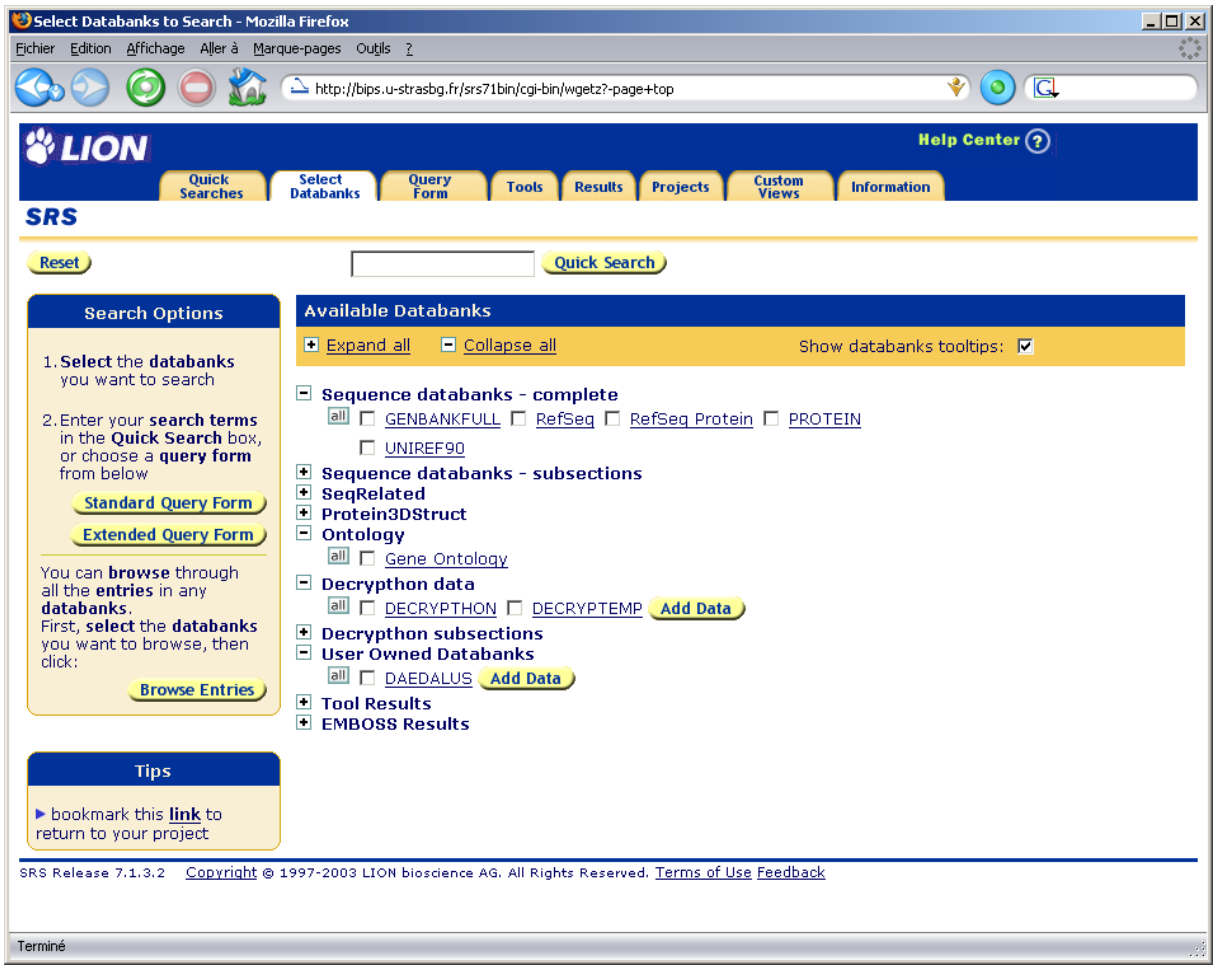


Figure 45. Screenshot of the SRS web server at the IGBMC.

5.4 PipeAlign protein family analysis toolkit

PipeAlign (Plewniak et al., 2003) is a protein family analysis toolkit developed in the laboratory. The pipeline integrates a six step process ranging from the search for sequence homologues in the protein and 3D structure databases to the definition of the hierarchical relationships within and between subfamilies.

The complete, automatic pipeline takes a single sequence or a set of sequences as input and performs an initial BlastP search in the UniProt and PDB databases. The database search is followed by a cascade of five different sequence analysis programs, shown in Figure 46 and described in detail below.

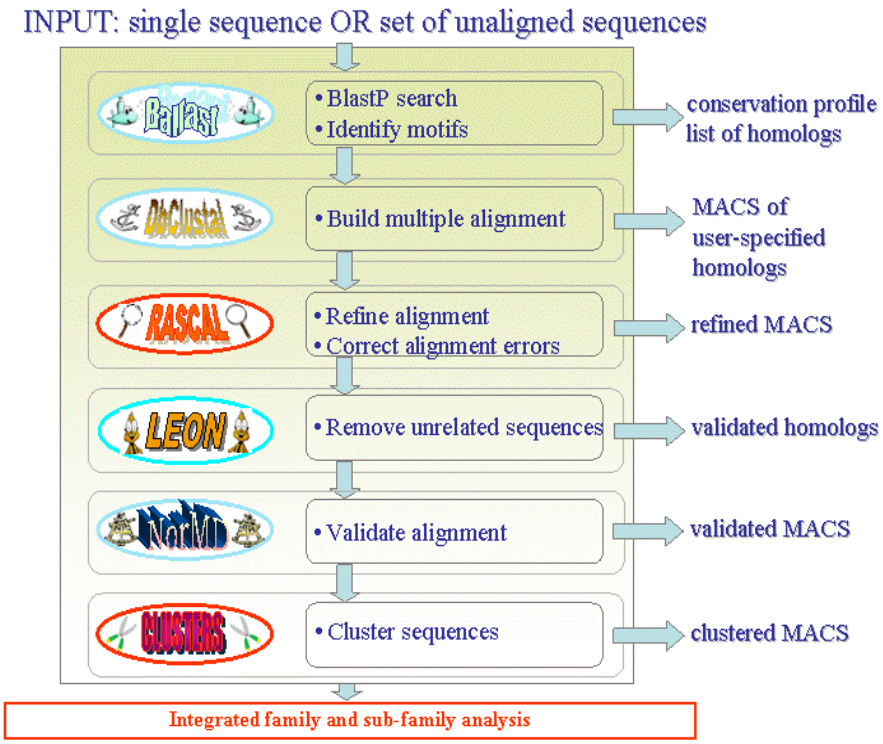


Figure 46. Overview of PipeAlign multiple alignment construction pipeline. (adapted from Plewniak *et al.* (Plewniak et al., 2003))

5.4.1 Ballast: post-processing of BlastP results

Ballast (Plewniak et al., 2000) builds a conservation profile of the database hits detected by BlastP. The contribution of each database hit is proportional to its significance, i.e. its E-value. The conservation profile is smoothed and then peaks are detected using the second derivative of the smoothed profile. These peaks define local maximum segments (LMSs) that correspond to sequence segments that are more conserved than their flanking regions. The positions of the LMSs in the query and database sequences are identified and are stored in a file as a list of anchors for input to DbClustal.

5.4.2 DbClustal: construction of the MACS

DbClustal (Thompson et al., 2000) integrates the local conservation information from the Ballast LMS or 'anchor' file in the global multiple sequence alignment program, ClustalW (Thompson et al., 1994). ClustalW incorporates the global dynamic programming algorithm developed by Needleman and Wunsch ((Needleman and Wunsch, 1970)). The recursive algorithm was modified in DbClustal, such that the score for aligning any pair of residues combines the residue comparison matrix score for the two amino acids and the anchor scores from the Ballast program. An anchor propagation is also incorporated, as Ballast determines

anchors for each database sequence relative to the query sequence only. Therefore, DbClustal propagates these anchors between all the sequences. The alignment weighting scheme implemented in DbClustal means that the global alignment is encouraged towards, but not constrained to, the conserved motifs.

5.4.3 RASCAL: rapid scanning and correction of alignment errors

DbClustal is a heuristic algorithm that can sometimes introduce errors into the multiple alignment. The RASCAL program (Thompson et al., 2003) is designed to detect these errors and to correct them. The multiple alignment output by DbClustal is divided horizontally and vertically to form a lattice in which well aligned, reliable regions can be differentiated. Potential errors are detected by comparing profiles of the reliable regions. RASCAL then performs a single re-alignment of each badly aligned region using an algorithm similar to that implemented in ClustalW (Thompson et al., 1994).Alignment correction is restricted to the less reliable regions only, leading to a more reliable and efficient refinement strategy.

5.4.4 LEON: removing of unrelated sequences

The next step in the pipeline is designed to detect the sequences in the MACS that are unrelated to the query sequence. The LEON program (Thompson et al., 2004a) uses the reliable regions, or 'core blocks' determined by RASCAL. Taking advantage of the transitive nature of homologous relationships, information from intermediate sequences is used to help define the conserved core blocks for the more divergent sequences. The conserved core blocks for each subfamily in the MACS are then chained together to form contiguous regions that are considered to be homologous to the query sequence. The amino acid composition of the sequences is also taken into account by the incorporation of a number of different algorithms for the detection of compositionally biased segments. Finally, any sequences that do not contain any homologous regions are removed from the MACS. The output from LEON is thus a high quality MACS containing only those sequences that share at least one homologous region with the query.

5.4.5 NorMD: MACS quality evaluation

The NorMD objective function (Thompson et al., 2001) is used to evaluate the quality of the MACS produced by the four previous steps. NorMD combines the advantages of a column-scoring technique with the sensitivity of methods incorporating residue similarity scores.

Here, a multiple alignment with a NorMD score greater than 0.3 is considered to be mostly well aligned.

5.4.6 Secator and DPC: sequence clustering

The Secator program (Wicker et al., 2002) clusters the sequences in a multiple alignment into potentially functional subgroups. The number of subgroups is determined automatically by the program.

The final output of the PipeAlign system is a high-quality, validated MACS, in which sequences are clustered into potential functional subgroups. PipeAlign can be used in local for high-throughput processing and is also available for interactive use via the web server at http://bips.u-strasbg.fr/PipeAlign/.

Chapter 6 - PromAn: an integrated knowledge-based tool dedicated to promoter analysis

A huge amount of data is now available, and a large number of programs have been developed for the prediction of promoters. However, most of these programs produce a large number of false positives. An efficient way to improve *in silico* promoter analysis, therefore, would be the combination of the results of several of the available promoter dedicated programs. Therefore, PromAn has been developed as an integrated web-based tool. It integrates distinct complementary databases, methods and programs dedicated to promoter analysis. The PromAn program provides automatic analysis of a genomic region with minimal prior knowledge of the genomic sequence.

Three versions of the PromAn program have been developed in response to distinct requirements: a local version, a web version and a high-throughput version. These different versions are described in detail below.

6.1 PromAn local version

6.1.1 Approach

The local version has been developed for dedicated promoter analysis in the context of collaboration projects with biologists. Therefore, the databases and programs integrated in PromAn had to exist in a downloadable version, compatible with our informatics system. The programs used in PromAn have been locally installed, compiled and modified for local use when necessary. Parsers have been developed for each database to format them for PromAn and to read the output files.

As PromAn is an integrated tool aimed at the combination of different programs, databases and methods to improve the efficiency of promoter analysis, it has been developed with a modular organization. Indeed, PromAn is a versatile and easily evolutive tool. A strict programming organization is at the basis of the upgradability of such an integrated tool. The modular organization of PromAn will be detailed in the subsequent sections.

PromAn generates a huge amount of data that need to be easily accessible. These data include results from several database and program requests, produced at each step of the PromAn program. They are analysed, filtered, extracted, formatted at each step of this program and finally the data of interest are integrated in the PromAn results file.

Therefore, dedicated functions have been developed associated with each file created and/or analysed by PromAn, which allow the automatic interrogation of all the available data. These functions are typically named "*Load…*" and "*Ask…*" (Figure 47).

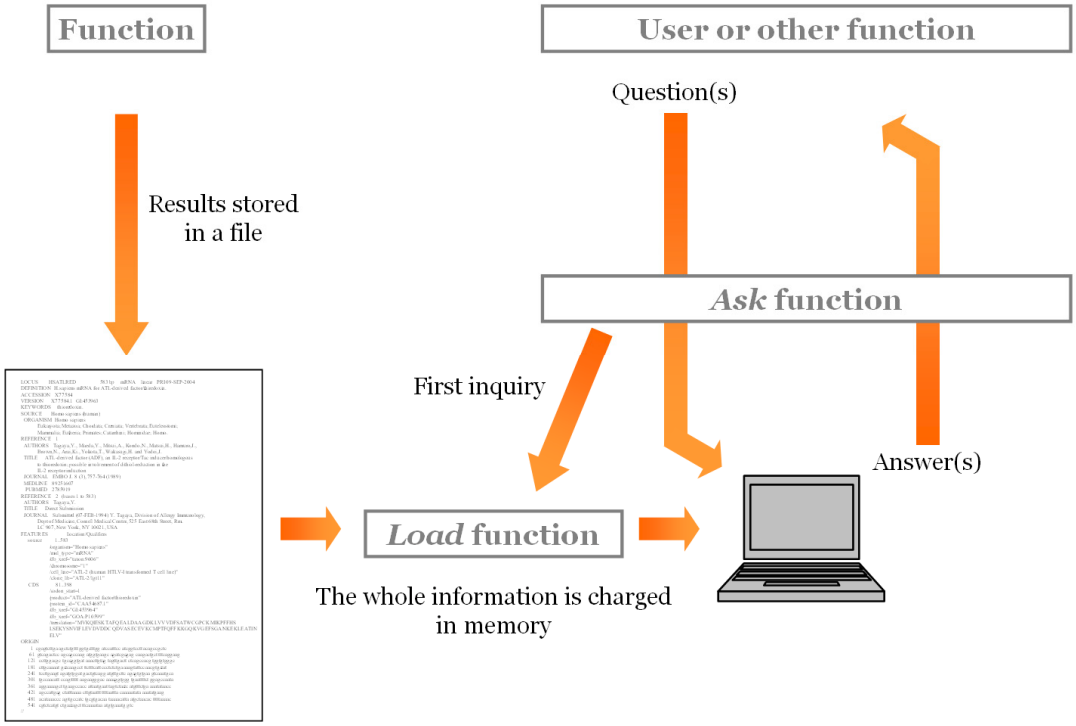


Figure 47. Creation, reading (*Load* function) and querying (*Ask* function) of files in the PromAn program.

Thus, LoadDBTSS allows the loading of all the data available in the DBTSS database and similarly LoadMatch allows the loading of the TFBSs predictions results provided by the match program. For the first query in a given file through an "*Ask…*" procedure, the complete data are read and indexed in a table available in random access memory using the corresponding "*Load…*" function. This indexation stored in memory is quickly accessible and allows as many subsequent requests in the file as necessary without needing to read this file again. Once the file use is over, the "*Ask…*" procedure allows the deletion of all the data indexed in the table. It is important to note that several files can be loaded in memory at any one time.

6.1.2 PromAn modules

The modular organization is at the basis of the versatility, efficiency and potential evolution of PromAn. The different modules implemented in PromAn are depicted in Figure 48.

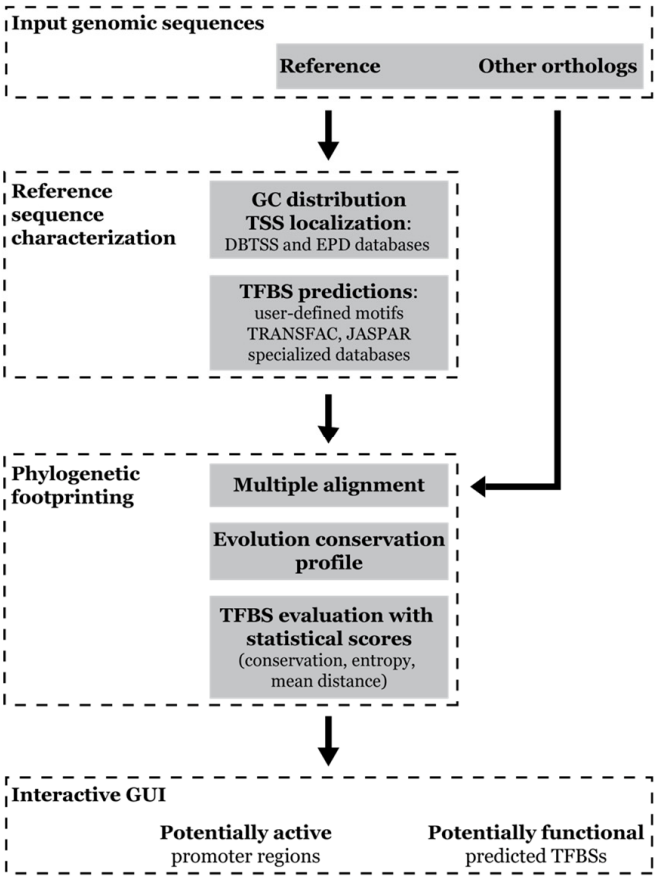


Figure 48. Flowchart of the PromAn integrated strategy.
The different modules that compose PromAn are depicted with a grey background. Briefly, a single reference sequence and a set of orthologous genomic sequences are required as input. First, dinucleotide distribution, TSS and promoter location as well as TFBS predictions are used to characterize the reference sequence. Next, a phylogenetic footprinting approach is used to determine the evolutionary conservation profile and TFBS evaluation with statistical scores. This integrated strategy is used to validate the TSS location, highlighting potentially active promoter regions and potentially functional TFBSs through the PromAn graphical user interface (GUI).

PromAn requires as input a single genomic sequence that is taken as the reference in subsequent analyses and a set of orthologous sequences in fasta format. First, dinucleotide distribution, TSS and promoter location as well as TFBS predictions are used to characterize the reference sequence. Next, a phylogenetic footprinting approach is used to determine the evolutionary conservation profile and TFBS evaluation with statistical scores.

This integrated strategy is used to validate the TSS location, highlighting potentially active promoter regions and potentially functional TFBSs through the PromAn Graphical User Interface (GUI).

6.1.2.1 Input genomic sequences

PromAn requires as input a single genomic sequence that will be taken as reference in subsequent analysis and a set of orthologous sequences in fasta format. The "orthologous promoter sequences" should contain the promoter sequences that are orthologous to the query promoter sequence. This set of orthologous sequences is used to evaluate the evolutionary conservation with respect to the reference sequence. PromAn provides the possibility of inputting large genomic sequences, which allows the user to begin analysis with minimal prior knowledge of proximal and distal active promoter regions.

Alu repetitive elements have been shown to house TFBSs (Compe et al., 2005). Suzuki *et al.* (Suzuki et al., 2004a) described that repeat sequences do not disturb block alignments (see 3.6.2). Thus, no tool has been implemented in PromAn for an eventual treatment of the repeats. The choice is left to the user to provide input sequences with or without previous repeat sequence masking. Masking of repeat sequences can be done by the user prior to PromAn using programs such as RepeatMasker (Smit, 1996) (http://www.repeatmasker.org/) to generate a masked sequence prior to alignment.

6.1.2.2 Reference sequence characterization

6.1.2.2.1 Dinucleotide profile

PromAn determines the nucleotide distribution of the reference sequence. Indeed, this distribution allows the detection of the presence of potential GC-rich regions within the given promoter of interest or a position where the nucleotide distribution switches from GC rich to AT rich or inversely. This allows the user to characterize the promoter according to the properties that can be assigned from the dinucleotide distribution (see 3.2.3).

6.1.2.2.2 TSS localization

In order to locate the promoter region on the reference genomic sequence and validate the TSS location, PromAn integrates experimentally-based databases and prediction programs.

The promoter databases DBTSS (Database of Transcriptional Start Sites) (Yamashita et al., 2006) and EPD (Eukaryotic Promoter Database) (Schmid et al., 2006) are integrated in PromAn. These databases are presented in section 4.2.1 . They allow either to confirm the

user presumed TSS, or to highlight a different position of the TSS, or to point out the presence of alternative TSSs.

If the promoter of the user input genomic sequence is not present in the experimental databases, prediction programs can help to determine the TSS position(s). TSSs and promoters prediction programs were briefly introduced and discussed in section 4.2.2 . Promoter, first exon or exonic map prediction programs (FirstEF, Eponine and GenScan) are integrated in PromAn.

- FirstEF

The FirstEF program (Davuluri et al., 2001) is dedicated to the task of identifying promoter regions and first exons in human and other mammalian genomes. It consists of different discriminant functions structured as a decision tree. These functions can recognize structural and compositional features such as CpG islands, promoter regions and first splice-donor sites.

The FirstEF algorithm can be described in several major steps. FirstEF scans the input sequence for potential first splice-donor sites (GT). Every GT dinucleotide is compared to a first exon database build by the author in order to determine the candidate splice-donor sites. For every candidate splice-donor site, FirstEF scans a region 2,000 bp in length (1,500 bp upstream and 500 bp downstream of GT) for the existence of a CpG window. FirstEF decides during this step whether the first exon is CpG-related or non-CpG-related. Then, FirstEF evaluates within the 1,500 bp region upstream of the candidate splice-donor site whether a sliding window of 570 bp length (500 bp upstream of TSS and 70 bp downstream of TSS) might be a promoter. Finally, FirstEF discriminates first exons from other genomic regions depending on whether or not the first intron region is GC rich and on whether or not the first exon region is CpG related. FirstEF is available at the following address: http://rulai.cshl.org/tools/FirstEF/ .

- Eponine

The Eponine program (Down and Hubbard, 2002) aims to predict the exact location of TSSs. Eponine models consist of a collection of positioned constraints, each represented by a DNA weight matrix (Bucher 1990). Eponine is based on model training. The eponine trainer combines a vector machine algorithm with a Monte Carlo sampling process. Training on a mouse cDNA data set gave the model shown in Figure 49.

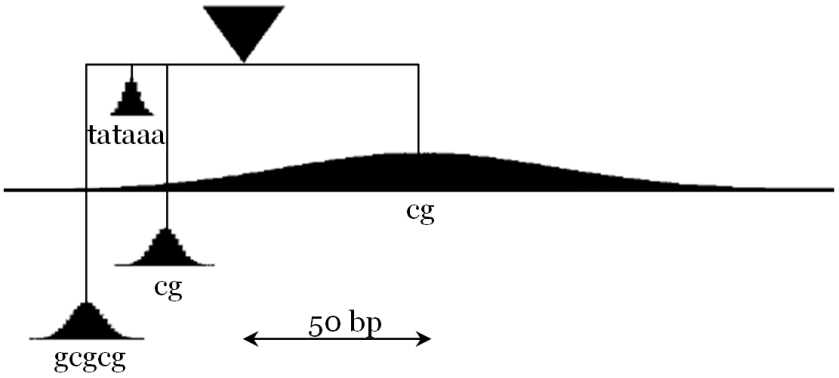


Figure 49. Schematic of Eponine core promoter model.
Extracted from (Down and Hubbard, 2002).

This model consists of four elements: (1) a diffuse preference for CpG enrichment downstream of the TSS; (2) a TATAAA motif, with a tightly focused distribution centered at position -30 relative to the TSS; (3 and 4) two GC-rich matrices closely flanking the TATA box.

- GenScan

The GenScan program (Burge and Karlin, 1997) aims at identifying complete exon/intron structures of genes in genomic DNA. This program has the capacity to predict multiple genes in a genomic sequence, deal with partial as well as complete genes, and to predict genes on both DNA strands. GenScan is particularly useful for the detection of genes recognized by the general transcriptional, splicing and transcriptional machinery which process most or all protein coding genes. Thus, GenScan is based on a probabilitic model which integrates TATA box and cap sites, present in most eukaryotic promoters. Furthermore, GenScan uses a Markov model of coding regions, donor and acceptor splice sites. The GenScan program is available at the following address: http://genes.mit.edu/GENSCAN.html .

6.1.2.2.3 TFBS predictions

Different methods have been implemented in PromAn to predict potentially active TFBSs. The user has the possibility to either input pre-defined TFBSs or to use the TFBSs described in several databases.

User-defined TFBSs can be used to look for specific TFBSs that have been described in the literature or experimentally characterized. These sequences can be given to PromAn either through consensus sequences or through Position Frequency Matrices (PFMs).

The consensus sequences should be in IUPAC nucleotide code (see Table 9) and in fasta format. An example of the consensus sequence format required is presented in Figure 50.

```
>TATA-box
TATAAA
>CAAT-box
CCAAT
```

Figure 50. Format of a user-defined TFBS pattern.

The PFMs are presented in the previous section 4.3.1 and are depicted in Figure 33. The PFMs should be described in the following order A, C, G and T with one line per nucleotide, the nucleotide beginning the corresponding line; and in fasta format. An example of the format required to describe the PFMs is provided in Figure 51.

```
>PFM_TATA-box
A  61  16  352  3  354  268  360  222  155  56  83  82  82  68  77
C  145 46  0    10 0    0    3    2    44   135 147 127 118 107 101
G  152 18  2    2  5    0    10   44   157  150 128 128 128 139 140
T  31  309 35   374 30  121  6    121  33   48  31  52  61  75  71
>PFM_CAAT-box
A  34  16  7    58 51   0    2    112  116  0   14  66  13  39  36  25
C  37  33  51   14 4    116  113  0    0    1   65  6   20  43  9   35
G  27  26  25   41 56   0    1    1    0    0   33  42  73  22  47  29
T  18  41  33   3  5    0    0    3    0    115 4   2   10  12  24  27
```

Figure 51. Format of a user-defined PFM.

The other possibility is to use the TFBS databases. The JASPAR and TRANSFAC databases have been implemented in PromAn.

The current JASPAR database is available in the PromAn tool. It has been presented in the section 4.3.3.1 . This database is freely available and downloadable. The user has the possibility to choose between the three databases available in JASPAR.

The Professional TRANSFAC database release 10.1 can also be used to predict TFBSs on the input reference sequence. As presented in the section 4.3.3.2 , the TRANSFAC database provides the opportunity to use different profiles that are available in PromAn for the TFBSs search.

TFBSs can also be predicted in PromAn from databases dedicated to a specific biological problem, such as retina or nuclear receptor databases. The consensus sequences of PFMs present in these databases were extracted from the literature.

The MATCH™ program is implemented in the PromAn tool in order to predict the TFBSs on the input reference sequence. This program has been previously described in the section 4.3.2 .

6.1.2.3 Phylogenetic footprinting

6.1.2.3.1 Multiple alignment

Several methods of multiple alignment have been implemented in the local version of PromAn.

At the very beginning of this development, the Clustal programs, which are the most popular global alignment programs, were used to perform this multiple alignment step. Using ClustalW, relatively short gaps perturbed the overall alignment. The low conservation regions disturbed the alignment so that, in many cases, exons were misaligned. Nevertheless, as described in the section 4.5.3 , ClustalW were primarily developed for aligning multiple protein sequences and thus do not provide optimized alignments with genomic sequences. Therefore, the BLASTZ local pairwise alignment tool combined with the DbClustal program was tested. As presented in the section 4.5.3 , BLASTZ is based on the Gapped BLAST algorithm that has been redesigned for the alignment of long genomic sequences. The local alignments created by BLASTZ were given as input to DbClustal in order to define anchors for the multiple alignment construction. However, the resulting multiple alignments were still not reliable.

Therefore, we tested multiple alignment tools dedicated to genomic sequences. The best program actually available for this task is the TBA program (Threaded Blockset Aligner) (Blanchette et al., 2004). As presented in the section 4.5.3 , the TBA program is a suite of independently executing programs which aligns sequences in the same order and the same sense (see Annexe 2 -). It is freely available at the web portal of the Laboratory of Webb Miller at the Penn State University Center for Comparative Genomics and Bioinformatics (http://www.bx.psu.edu/miller_lab/).

The TBA program has been implemented in the PromAn tool. TBA is composed of three steps which are presented in Figure 52.

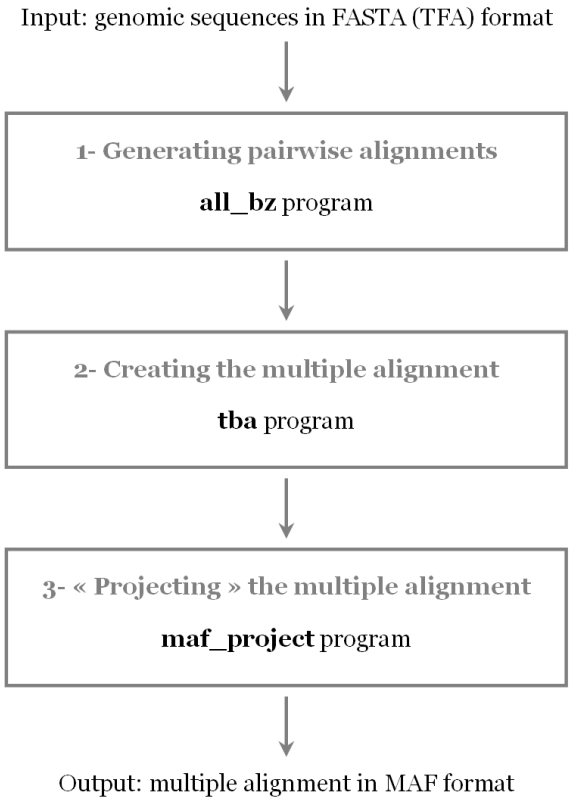


Figure 52. TBA algorithm steps.

The all_bz program generates pairwise alignments with the BLASTZ program for all pairs of specified sequences. The resulting pairwise alignment files are in MAF (Multiple Alignment Format) format. This format is presented in Figure 53.

```
##maf version=1 scoring=tba.v8
# tba.v8(((human chimp) baboon) (mouse rat))
# multiz.v7
# maf_project.v5 _tba_right.maf3 mouse _tba_C
# single_cov2.v4 single_cov2 /dev/stdin

a score=23262.0
s hg16.chr7     27578828 38 + 158545518 AAA-GGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s panTro1.chr6 28741140 38 + 161576975 AAA-GGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s baboon          116834 38 +   4622798 AAA-GGGAATGTTAACCAAATGA---GTTGTCTCTTATGGTG
s mm4.chr6      53215344 38 + 151104725 -AATGGGAATGTTAAGCAAACGA---ATTGTCTCTCAGTGTG
s rn3.chr4      81344243 40 + 187371129 -AA-GGGGATGCTAAGCCAATGAGTTGTTGTCTCTCAATGTG

a score=5062.0
s hg16.chr7     27699739  6 + 158545518 TAAAGA
s panTro1.chr6 28862317  6 + 161576975 TAAAGA
s baboon          241163  6 +   4622798 TAAAGA
s mm4.chr6      53303881  6 + 151104725 TAAAGA
s rn3.chr4      81444246  6 + 187371129 TAAAGA

a score=6636.0
s hg16.chr7     27707221 13 + 158545518 GCAGCTGAAAAACA
s panTro1.chr6 28869787 13 + 161576975 GCAGCTGAAAAACA
s baboon          249182 13 +   4622798 GCAGCTGAAAAACA
s mm4.chr6      53310102 13 + 151104725 ACAGCTGAAAATA
```

Figure 53. MAF, Multiple Alignment Format.
The MAF stores a series of multiple alignments. The first line beginning with ## is the header line. Lines starting with # are alignment parameter lines. Each multiple alignment block is in a separate paragraph that begins with an "a" line and contains an "s" line for each sequence in the multiple alignment. Lines starting with "s" represent a sequence within an alignment block. This line contains the name of the source sequence, the start position of the aligning region, the size of the alignment, the strand (if "-", then the alignment is to the reverse-complemented source), the size of the entire source sequence (not just the part involved in the alignment) and finally the nucleotides and insertions in the alignment.

The all_bz program requires an evolutionary tree as input argument. This must be a binary tree described in a modified Newick format (see http://evolution.genetics.washington.edu/phylip/newicktree.html for details) where the branch lengths are removed, spaces are substituted for commas, and the final semi-colon is removed. This specied-guid-tree consists of double quotes, parenthesis, and species names, e.g. (((((((human chimp) gorilla) baboon) (rat mouse)) (cow pig)) chicken) fugu). The version of a tree used by Webb Miller is depicted in Figure 54.

```
((((((((((human chimp) gorilla) orangutan) ((baboon macaque) vervet))
((marmoset dusky_titi) squirrel_monkey)) ((mouse_lemur lemur) galago))
((rat mouse) rabbit)) (((((((cow sheep) muntjak_indian) pig) ((cat dog)
horse)) (ajbat cpbat)) hedgehog) armadillo)) (((opossum monodelphis)
wallaby) dunnart)) platypus) (chicken tortoise)) ((tetra fugu) zfish))
```

Figure 54. Phylogenetic tree between 37 species in Newick format.

The TBA program generates the multiple alignment from the pairwise alignments generated with the BLASTZ program. It takes as input the same phylogenetic tree as the all_bz program. The input genomic sequences file should be named with the same names used in the phylogenetic trees.

Finally, the maf_project program "projects" the multiple alignment to a given input sequence defined as the reference sequence. The resulting maf blocks always have the reference sequence in the top row, and the maf blocks are ordered by the starting position of the top row. Thus, the final output is a multiple alignment file in maf format and projected onto a reference sequence. Indeed, such an alignment describes the blocks of sequences that align with the reference sequence.

6.1.2.3.2 Conservation profile

The conservation profile (*n*) is calculated relative to the reference sequence and for each position of the multiple alignment. At each position, it corresponds to the number of sequences having the same nucleotide as the reference sequence. In other words, it corresponds to the number of sequences in which the nucleotide of the reference sequence is conserved. This calculation is explained in the Figure 55.

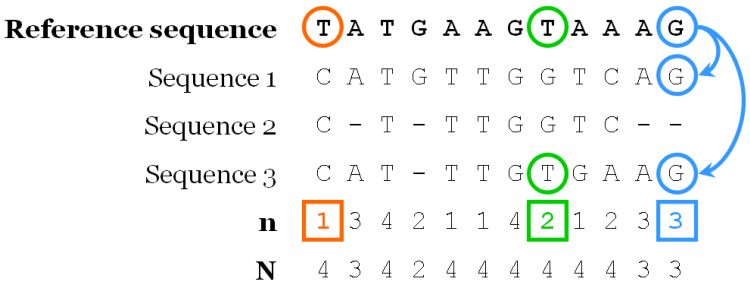


Figure 55. n and N calculation.

This conservation profile is named the "*n* profile" where "*n*" is the number of residues identical to the reference sequence residue at a position of the promoter multiple alignment.

Other profiles are also calculated. The "*N* profile" is a profile where "*N*" is the number of sequences that are present at a position of the promoter multiple alignment. The "*n*/*N* profile" is a profile with the ratio "*n*/*N*" corresponding to *n* divided by *N* at each position of the multiple alignment. This profile is particularly informative in the context of multiple alignments in MSF format. In fact, with the method using the TBA program, only regions containing sequences sharing a high homology are aligned. The other regions are not aligned at all. So, in this case the *n*/*N* is either equal to "0" in regions that are not conserved or *n*/*N* is close to "1" in conserved regions. In the case of a global multiple alignment each residue of all the sequences is present in the alignment, so gaps are present at the positions that do not align. In this situation, the degree of conservation, reflected by "*n*/*N*", will be highly variable all along the promoter multiple alignment and will thus be very informative.

6.1.2.3.3 TFBSs evaluation and scoring

Three methods of scoring of the predicted TFBSs have been implemented in PromAn : a conservation score, an entropy score and a mean distance score.

- Conservation score

The conservation score measures the identity of the orthologous sequences with respect to the reference sequence for a given region. This conservation score is the average value of the ratio "*n*/*N*" (see Figure 55) calculated on the predicted TFBS region defined by its start position (BR_{seq}) and end (ER_{seq}) positions on the Reference sequence (R_{seq}).

$$Conservation\,Score = \frac{\sum_{i=BR_{seq}}^{ER_{seq}} (\frac{n_i}{N_i})}{(ER_{seq} - BR_{seq}) + 1}$$

If only the Reference sequence is present at a given position, *n* and *N* are equal to 1 and the ratio *n*/*N* is set to zero.

- Entropy score

The entropy score is based on the ScoreCons program (Valdar, 2002). It measures the degree of nucleic acid variability to quantify residue conservation in a multiple alignment. The Entropy score used in the ScoreCons program normalizes Shannon's entropy.

Shannon's entropy is given by: $S = -\sum_{a}^{K} p_a \times \log_2 p_a$

With *N* = number of residues in a column

K = number of residue types (A C T G -), thus *K* = 5

n_a = number of residues of type a

$$p_a = \frac{n_a}{N}$$

Normalized Shannon's Entropy is given by: $Centropy = \frac{-\sum_{a}^{K} p_a \times \log_2 p_a}{\log_2(\min(N, K))}$

Thus, the Entropy score for a TFBS site is given by:

$$Entropy\,Score = 1 - \frac{\sum_{i=BRseq}^{ERseq} Centropy_i}{(ERseq - BRseq) + 1}$$

This score is equal to 1 when the column is totally conserved (low entropy) and to 0 when the column is diverse (high entropy).

- Mean distance score

The mean distance score based on the ClustalX conservation profile (Thompson et al., 1997), corresponds to the mean pairwise distance between sequences in a continuous sequence space using the BLASTZ nucleotide substitution matrix (Schwartz et al., 2003) (Table 18).

	A	C	G	T
A	0.91	-1.14	-0.31	-1.23
C	-1.14	1.00	-1.25	-0.31
G	-0.31	-1.25	-1.00	-1.14
T	-1.23	-0.31	-1.14	0.91

Table 18. Nucleotide substitution matrix used in the BLASTZ program

The mean distance at each position (j) of the multiple alignment (MD_j) is defined as:

$$MD_j = e^{-\frac{D}{N^2}} \times \frac{N}{M}$$

With D, calculated on all possible pairs of nucleotides (NA1,NA2) given by:

$$D_j = \sum_{NA1,NA2} n_j \times m_j \times d$$

With *nj*, the number of nucleotides *NA1* at the jth position of the multiple alignment. m_j, the number of nucleotides *NA2* at the jth position of the multiple alignment. And d, the substitution distance between the nucleotide *NA1* and *NA2* according to the nucleotide substitution matrix depicted in Table 18.

The score of mean distance for a TFBS predicted between the position BR_{seq} and ER_{seq} is given by:

$$MD\,Score = \frac{\sum_{i=BR_{seq}}^{ER_{seq}} MD_i}{(ER_{seq} - BR_{seq}) + 1}$$

These three conservation calculation methods are maintained in PromAn as they have been observed to provide different results. Thus, it would be interesting to study these scores to highlight the benefit of each method and to be able to combine these scores to estimate the conservation.

.

6.1.2.4 PromAn Graphical User Interface (GUI)

The PromAn GUI is an independent module allowing the visualization and exploration of the results. The PromAn GUI allows the integration of all the results in a single interface to allow the user to easily compare and cross-validate the results. The PromAn GUI thus helps the biologist to refine results and further guide gene regulation hypotheses in parallel with experimental data and validations.This GUI is implemented in the Tcl/Tk language. The PromAn GUI will be further detailed in the example of the rhodopsin gene in the Chapter 9 - of the "Results and Discussion" section.

The GUI allows a visualization of all the results simultaneously relative to the reference sequence. The user can choose the results to be displayed in the interface. The conservation profiles can be depicted with respect to the reference sequence. The user can zoom in on the reference sequence at the nucleotide level. All the profiles are also available in a "smoothed" version. Indeed, they are "smoothed" (each value is averaged on a sliding windows of 10 bp) in order to be more informative at the scale of the whole promoter sequence or of a large region.

The nucleotide distribution profiles can also be depicted in the same graph as the conservation profile in order to be able to integrate all the results. Both AT and GC dinucleotide distribution profiles are available although this information is redundant. Thus, these symmetrical curves allow a rapid visualization of the dinucleotide distribution and a rapid analyse of its evolution along the reference sequence. The interpretation of the dinucleotide distribution is not so intuitive if a single profile is depicted. Nevertheless, the user can select the profiles to be displayed.

The user can also filter and select the TSS validation and prediction results to be displayed. The TFBS predictions for display can be filtered according to: the initial database or matrix profile; their identifier; or the different processed scores (matrix score, core score, conservation score). The reference sequence, the conservations profiles and the TFBS predictions are juxtaposed relative to the nucleotide positions of the reference sequence.

The use of the PromAn GUI is detailed at the following address: http://bips.u-strasbg.fr/PromAn/PromAnGUIHelp.html.

6.1.3 Output results

As output PromAn provides a results file (with the extension « .PromAn ») corresponding to the input file of the PromAnGUI application. Moreover the genomic sequence multiple

alignment in maf and fasta formats are also supplied. All the intermediate files can be provided if required.

6.2 PromAn web version

6.2.1 Approach

In order to make PromAn freely accessible to the scientific community, a web server has been developed, although the web version is an adapted version.

Indeed, all the programs and databases for the web server have to be freely available. This necessitates for example the use of the public version of the Transfac database (release 7.0). Moreover, a web server access implies limitations in terms of sequence size in order to limit the processing time. As a consequence, the input genomic sequences are limited to a maximal length of 20 kb. This size can be processed in a reasonable time, but still allows for variation on the TSS position. According to the NCBI statistics of the human genome, the average size of exons and introns are 231 and 5407 bp, respectively. Thus, large sequences up to 20 kb can allow for 5' non-coding exons, TSS mis-location and coding exons anchoring the multiple alignment.

The web server version of PromAn is available at http://bips.u-strasbg.fr/PromAn/ and has been published in the annual NAR Web Server Issue 2006 (Lardenois et al., 2006).

In the subsequent sections, only the implementations differing from the local version of PromAn are presented.

6.2.2 Implementation

6.2.2.1 TFBS predictions

The same possibilities for TFBS predictions as in the local PromAn version are available. The only difference is in the version of the TRANSFAC database. Indeed, the public TRANSFAC database (version 7.4) is used in the web verion of the PromAn program. This database is the most widely used database in which detailed information concerning TF-binding sites, which have been characterized by various experimental methods, is compiled (Kel et al., 2003).

6.2.2.2 Multiple alignment

The TBA program is not currently implemented in the web server version of PromAn so that the user does not have to provide a phylogenetic tree in Newik format as input. The multiple alignment step was implemented as described in Figure 56.

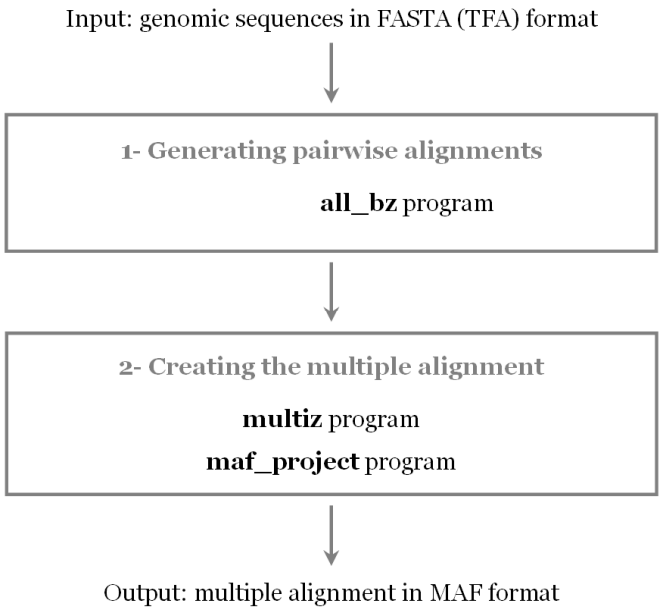


Figure 56. Multiple alignment processing implemented in the web server version of PromAn.

The *all_bz* program is first used to generate the pairwise alignments for each pair of sequences with the BLASTZ program. The second step allows the construction of the multiple alignment from the previously created pairwise files. The MULTIZ program aligns the blocksets generated with BLASTZ. The *maf_project* program projects the maf file onto the reference sequence. The resulting maf blocks always have the reference sequence in the top row, and the maf blocks are ordered by the starting position of the top row.

6.2.3 Output results

Direct internet links to the output files of the PromAn web version are provided by e-mail. They include the PromAn results file, the PromAnGUI stand-alone application and the multiple alignment file in fasta format. This allows one to easily visualize, analyse and refine the results as often as needed in parallel with expert biological knowledge and experimental validations, which are indispensable to complete and further guide gene regulation hypotheses.

6.2.4 System requirements

The PromAn GUI application was written in Tcl/Tk which offers the possibility to run it on a wide variety of platforms, including Windows, Mac, and essentially all flavors of Unix (Linux, Solaris, etc.). PromAn GUI was developed and extensively used on a PENTIUM IV desktop computer with 1Gb of memory.

6.3 High-throughput version

6.3.1 Approach

The ideas described in section 4.6 imply a high-throughput promoter analysis. It is clear that the use of the local version of PromAn in high-throughput would not be possible because the calculation of the conservation scores from the multiple alignment for each predicted TFBSs take too much time. As a consequence, a high-throughput version of PromAn was developed.

6.3.2 Implementation

6.3.2.1 Automated genomic sequence extraction

For this high-throughput study, we have chosen to consider the mouse genome as a reference. To extract the genomic sequences, several inputs are allowed. The only constraint is that it should be possible to convert it into a mouse EntrezGene ID (Figure 57).

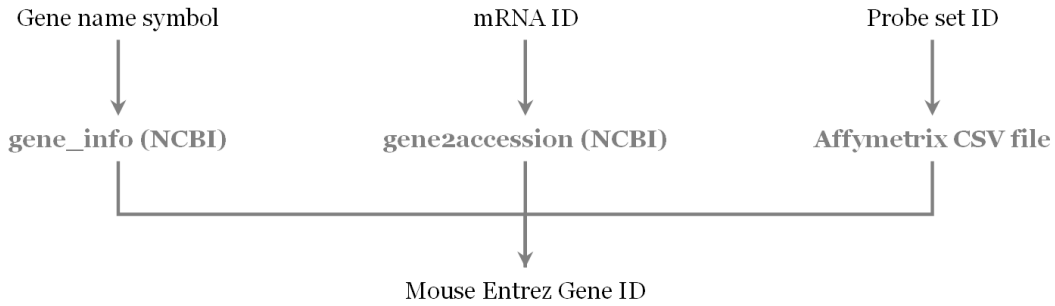


Figure 57. Conversion of input entries into EntreGene IDs.
The files allowing the conversion of the input into an Entrez Gene ID are depicted in grey. The conversion of the input gene name symbol, mRNA ID and probe set ID into a mouse Entrez Gene ID are done with the gene_info file (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_info.gz) from the NCBI, the gene2accession (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2accession.gz) file from the NCBI and the affymetrix CSV files respectively.

Entrez Gene (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) is NCBI's database for gene-specific information (Maglott et al., 2005). It focuses on the genomes that have been completely sequenced. Entrez gene provides unique identifiers (ID) for genes of model organisms, the mouse Entrez Gene IDs are considered in our analysis.

In the literature, genes are mainly described with a gene name. Such a denomination is dangerous because several symbols are used to define a single gene. For high-throughput studies, it was possible to identify the corresponding Entrez Gene ID if the gene name corresponding to the official gene name symbol. Thus, the *gene_info* file provided by the NCBI allows a conversion between the EntrezGene IDs and the gene names.

The following step aims at identifying the mouse mRNA IDs corresponding to each Entrez Gene ID identified and to subsequently locate their first exon on the mouse genome (Figure 58).

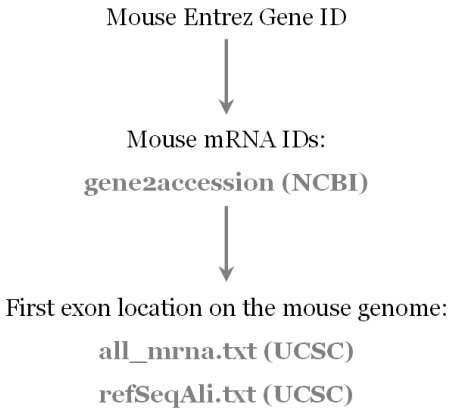


Figure 58. From the mouse Entrez Gene IDs to first exon localization on the mouse genome.

Briefly, the potential TSS of each gene was automatically extracted according to the following approach:

- The NCBI gene2accession file allows one to extract all the mRNA IDs of a given Entrez Gene ID (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2accession.gz).
- All the corresponding mRNA sequences are mapped to the mouse genome using the UCSC mapping files (all_mrna.txt and refSeqAli.txt).
- The first exon localization and by extension the TSS position of each mRNA is then determined.

All the determined TSS positions of a given Entrez Gene ID are further processed by the "first exon mapping" method to select a single TSS position per gene (Figure 59).

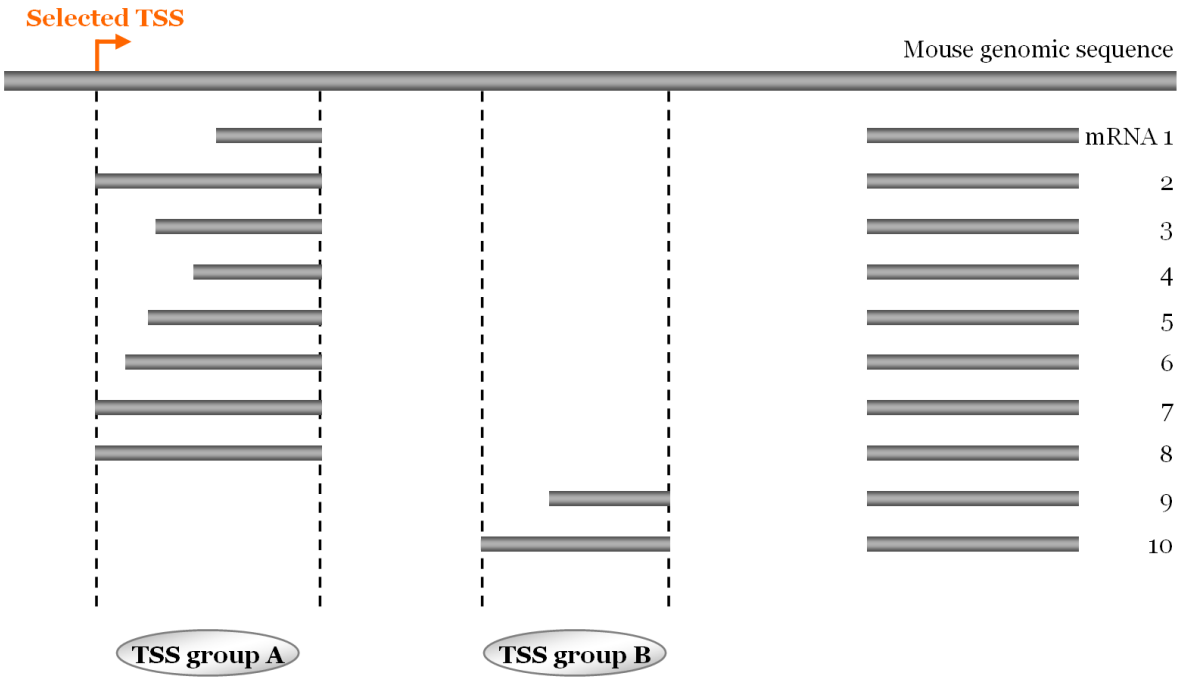


Figure 59. TSS determination for high-throughput analysis.
The first exons of ten mouse mRNAs corresponding to a single mouse Entrez Gene ID are mapped on the corresponding mouse genomic sequence. The First exons are grouped according to their overlapping genomic location ("TSS group A" and "TSS group B"). Only the TSS group containing the highest number of first exons ("TSS group A" in this example) is retained. Finally, the most 5' TSS of the selected group is considered as the reference or potential TSS for this Entrez Gene ID ("Selected TSS", depicted in orange).

This method allows the determination of a single "selected TSS" for each given mouse Entrez Gene ID. Then, mouse genomic sequences are extracted from – 10 kb to + 10 kb relative to the "selected TSS". Indeed, each input genomic sequence is 20,001 bp long. The program twoBitToFa from the blat suite of tools (http://hgwdev.cse.ucsc.edu/~kent/exe/) (Kent, 2002) is used to extract genomic sequence from the mouse genome file (2bit format). The 2bit format is a dense and quickly searchable format. This strategy allows a rapid extraction of the sequences on a complete eukaryotic genome. The mm8 version of the mouse genome and the reported data were obtained from the Build 36 "essentially complete" assembly by NCBI and the Mouse Genome Sequencing Consortium.

6.3.2.2 Conservation estimation

As the calculation of the scores evaluating the evolutionary conservation in PromAn is time-consuming for high-throughput analysis, pre-processed conservation scores were used. Indeed, the UCSC Genome Browser provides multiple genome alignments as well as the corresponding phastCons (Siepel et al., 2005) conservation scores. These scores are a measure

of evolutionary conservation in 17 vertebrates (including mammalian, amphibian, bird, and fish species) and are based on a phylogenetic Hidden Markov Model (HMM). These scores were calculated on the *multiz* multiple alignment of the mouse genome (mm8, Feb. 2006) with16 vertebrates genomes (Table 19).

Vertebrate organism	Genome assembly version
Mouse	Feb 2006, mm8
Rat	Nov 2004, rn4
Rabbit	May 2005, oryCun1
Human	Mar. 2006, hg18
Chimp	Mar. 2006, panTro2
Macaque	Jan 2006, rheMac2
Dog	May 2005, canFam2
Cow	Mar 2005, bosTau2
Armadillo	May 2005, dasNov1
Elephant	May 2005, loxAfr1
Tenrec	Jul 2005, echTel1
Opossum	Jan 2006, monDom4
Chicken	Feb 2004, galGal2
Frog	Aug 2005, xenTro2
Zebrafish	Mar 2006, danRer4
Tetraodon	Feb 2004, tetNig1
Fugu	Aug 2002, fr1

Table 19. Assemblies of the 17 vertebrate genomes aligned.

The phylogenetic tree used is based on Murphy *et al.* (Venter et al., 2001).

For this high-throughput promoter analysis with PromAn, only the mouse reference sequences need to be extracted as the multiple alignment and the conservation scores are pre-processed by the UCSC genome browser. For each mouse extracted genomic sequence, a file containing a phastCons score per position is extracted. Therefore, as previously described, the conservation score of each TFBS prediction corresponds to the average value of all the phastCons scores of the corresponding genomic region.

6.3.2.3 Promoter analysis

The promoter analysis is performed with the local version of PromAn (see section 6.1) with the exception of the multiple alignment and of the score calculations. The TFBSs are searched with the Professional version of TRANSFAC as well as with dedicated databases, for which the TFBSs have been extracted from the literature according to the biological projects.

Chapter 7 - Protein sequence annotation with GOAnno

The *Gene Ontology* (GO) (Ashburner et al., 2000) was originally constructed in 1998. Its main objective as its name indicates is the management of information related to genes. The GO project is a collaborative effort that addresses the need for consistent descriptions of gene products in different databases. The project began as a collaboration between three model organism databases, FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Database (MGD). Since then, the GO Consortium has grown to include many databases, including several of the world's major repositories for plant, animal and microbial genomes. The GO has become a standard tool in the bioinformatics arsenal.

7.1 Gene ontology presentation

The Gene Ontology is a controlled vocabulary of terms split into three related ontologies covering basic areas of molecular biology: the molecular function of gene products, their role in multi-step biological processes, and their localization to cellular components.

Each GO term consists of a unique alpha-numerical identifier, a term name, and a textual definition. Each term is also assigned to one of the three ontologies molecular function, biological process or cellular component.

- Molecular function describes activities, such as catalytic or binding activities, at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products.
- A biological process is a sery of events accomplished by one or more ordered assemblies of molecular functions. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct step. A biological process is not equivalent to a pathway.

- A cellular component is just that, a component of a cell, but with the proviso that it is part of some larger object; this may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).

The ontology file is freely available from the GO website (http://www.geneontology.org/); the terms can be searched and browsed online using the GO browser AmiGO (http://www.genedb.org/amigo/perl/go.cgi).

The controlled vocabularies are structured so that they can be queried at different levels. Thus, GO can be used to find all the gene products in the mouse genome that are involved in signal transduction. The ontologies are structured as directed acyclic graphs, which are similar to hierarchies but differ in that a child, or more specialized, term can have many parents, or less specialized, terms. Indeed, each gene product can be linked to a given level of GO according to what is known about this molecule. This characteristic offers a great flexibility allowing more or less specialized annotation Figure 60.

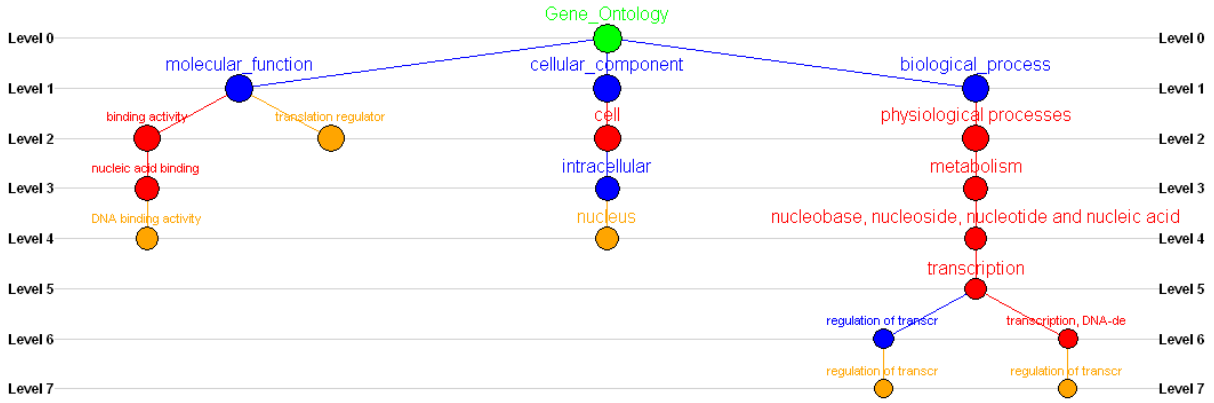


Figure 60. Example of a protein annotation with GO.
A gene product can have different functions, cellular localizations or cen be implied in multiple biological processes.

As an example, a well known protein will be assigne several specialized GO terms. A less well characterized protein will be assigned general GO terms such as "metabolism". Finally, a predicted gene product with unknown function will not be annotated with GO terms.

7.2 Approach

In the context of coexpressed gene annotation, to increase the quality and quantity of GO annotation, I have participated in the elaboration of a new method of propagation of the GO annotation, GOAnno (Chalmel et al., 2005) (http://bips.u-strasbg.fr/GOAnno) (see Annexe 1 -

). GOAnno, based on the study of functional protein sub-families allows to annotate unknown proteins or to complete the protein annotation in terms of GeneOntology.

Several methods had been developed to annotate proteins in GO. They generally employ sequence similarities selected based on best BLAST (Altschul et al., 1997) hits(Hennig et al., 2003) (Khan et al., 2003) (Zehetner, 2003) or a predefined subset of GO terms (Jensen et al., 2003). Nevertheless, the identification of orthologous sequences through the detection of the first hits in a blast request has often been shown to be inappropriate (Frickey and Lupas, 2004) (Koski and Golding, 2001) (Wall et al., 2003). As a consequence, the GOAnno method has been developed based on Multiple Alignments of Complete Sequences (MACS) (Lecompte et al., 2001). Furthermore, several methods require the user to define a relevant GO level (in the GO tree) whatever the branch considered while GOAnno automatically select the more relevant level of annotation that can be assigned to the protein. Indeed, the same level in different branches does not mean the same degree of specificity of the GO information.

7.3 Implementation

GOAnno allows the GO annotation of unkwown protein sequences. GOAnno takes advantage of the evolutionary information available in Multiple Alignments of Complete Sequences (MACS) (Lecompte et al., 2001) organized hierarchically into functional subfamilies. The members within subfamilies are conserved enough to filter, enrich and propagate GO terms using the GOAnno algorithm. Another originality is the absence of any predefined parameters such as GO level or subsets of GO terms. The flowchart of the GOAnno method is presented in Figure 61.

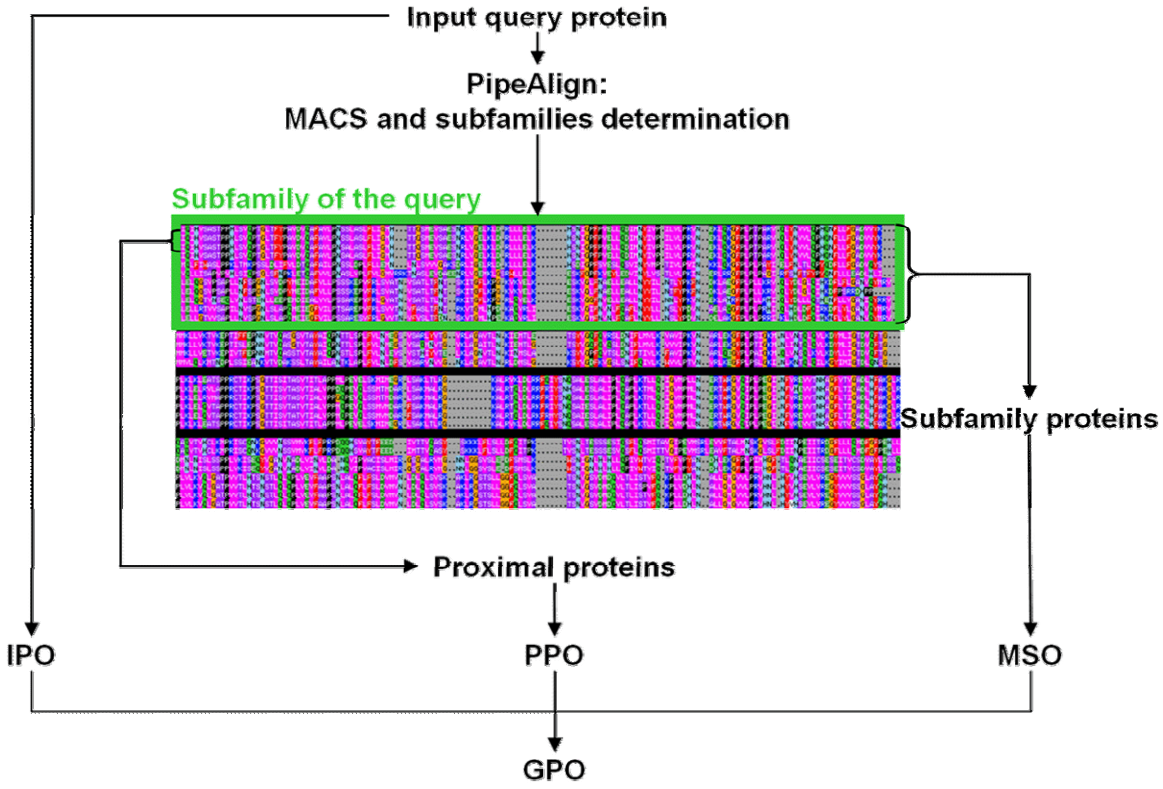


Figure 61. GOAnno algorithm flowchart.

7.3.1 Determination of the functional subfamily of the query protein using a MACS

This preliminary step incorporates the strategy used in PipeAlign, a toolkit for protein family analysis (see section 5.4). The cascade of the five PipeAlign analysis programs yields a hierarchised MACS of protein homologues clustered into potential functional subfamilies (Figure 62).

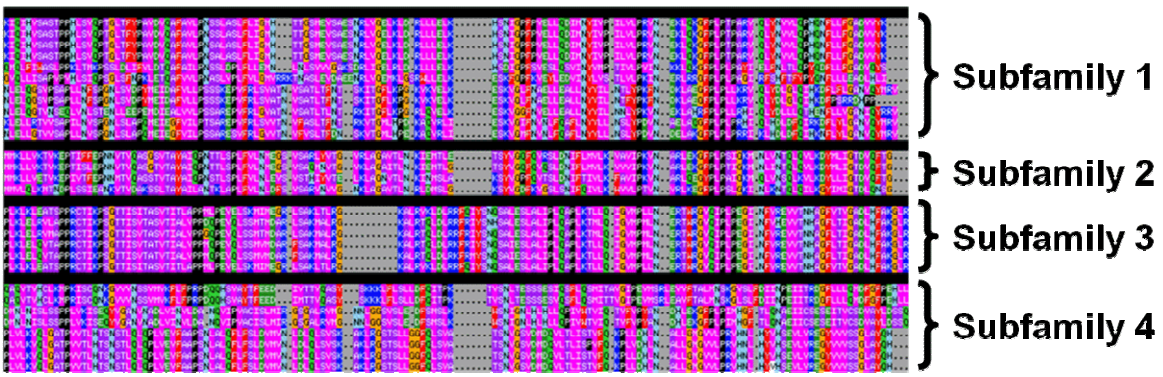


Figure 62. MACS clustered in potential functional subfamilies

Before presenting the GOAnno propagation algorithm, some vocabulary needs to be defined in the MACS (Figure 63).

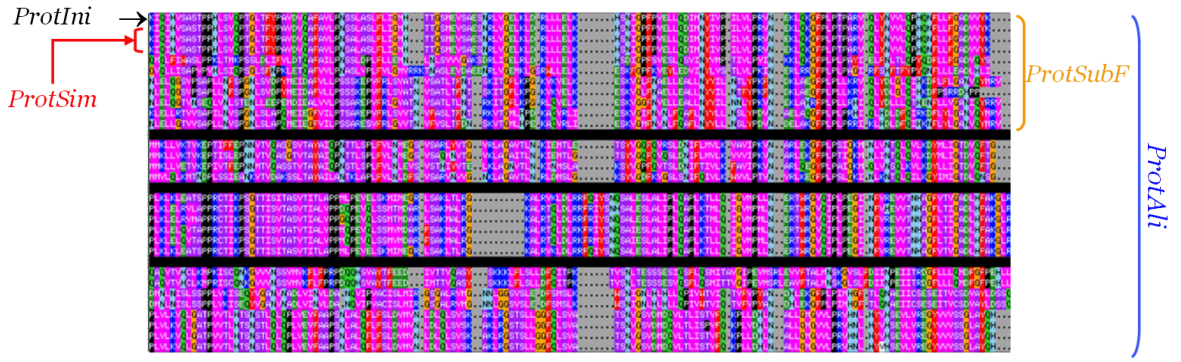


Figure 63. Definition of ProtIni, ProtSim, ProtSubF, and ProtAli in a MACS.
ProtIni, initial protein. *ProtSim*, proteins sharing more than 98% identity with the *ProtIni*. *ProtSubF*, proteins belonging to the same functional subfamily as the *ProtIni*. *ProtAli*, set of proteins present in the MACS.

ProtAli defines all the proteins present in the MACS, and *ProtF*, the proteins belonging to the functional subfamily containing the query protein (*ProtIni*). Finally, *ProtSim*, corresponds to all the proteins present in the MACS that share more than 98% of identity with the *ProtIni*.

To propagate GO information to the *ProtIni* based on the MACS, the *ProtPop* should first be defined as the population of proteins that will be involved in the propagation. *ProtPop* contains *ProtSim* and *ProtSubF* if the objective function NorMD for the query protein subfamily is higher than 0.3 (see section 5.4.5).

7.3.2 Major steps of the GOAnno algorithm

The major steps in the method implemented in the GOAnno program are described in detail. This method can be independently applied to each ontology: the molecular function, biological process or cellular compartment. The GOAnno algorithm can be described in four major steps. At the end of each step, the duplicated GO terms as well as the parent terms (less specialized on the same branch) are systematically removed (Figure 64).

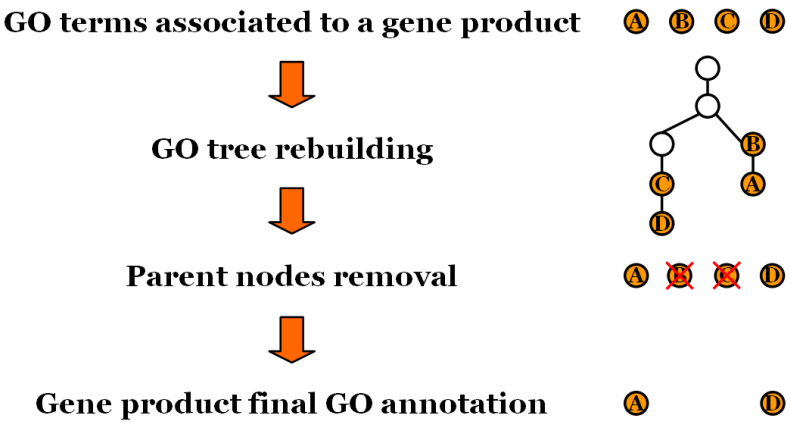


Figure 64. GO parent terms removal.

The first step is the extraction of the initial GO annotation for the considered gene product *ProtIni*. This annotation is named IPO *(Initial Protein gene Ontology)* (Figure 65).

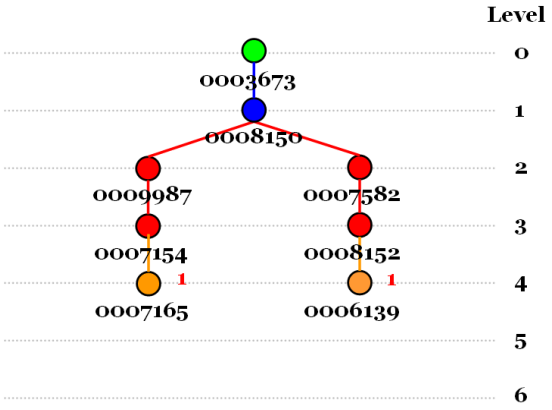


Figure 65. Example of a GO tree constructed from the IPO terms of a protein.
The IPO is defined with the most specialized GO terms, depicted here in orange. The root of the tree is the first level depicted in orange, it is called "Gene Ontology"

The IPO is defined by the whole GO terms deduced from conversion tables available at the GO Consortium (InterPro, Pfam, Prints, PRODOM, Prosite, SMART protein motifs, Enzyme Commission numbers and SWISS-PROT keywords to GO nodes) added to the native GO annotation available in the UniProt database (Figure 66).

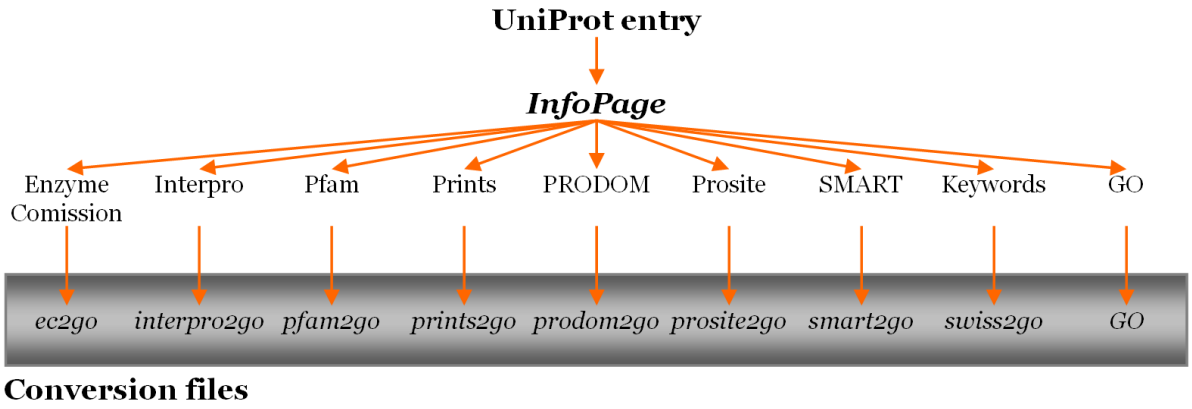


Figure 66. UniProt protein annotation with the GO.

The second step is the determination of the PPO (*Proximal Protein gene Ontology*). It represents the GO of proximal proteins (*ProtSim*) of the *ProtIni*. The set of the IPO terms of each *ProtSim* constitutes the PPO.

The third step is the determination of the MSO (*Mean Subfamily gene Ontology*) from the original propagation algorithm. The MSO is thus the GO terms can that reasonably be propagated to the whole population of proteins of the subfamily, *ProtPop*. The IPO terms of the *ProtPop* are collected and used to construct the corresponding GO tree (Figure 67 a).

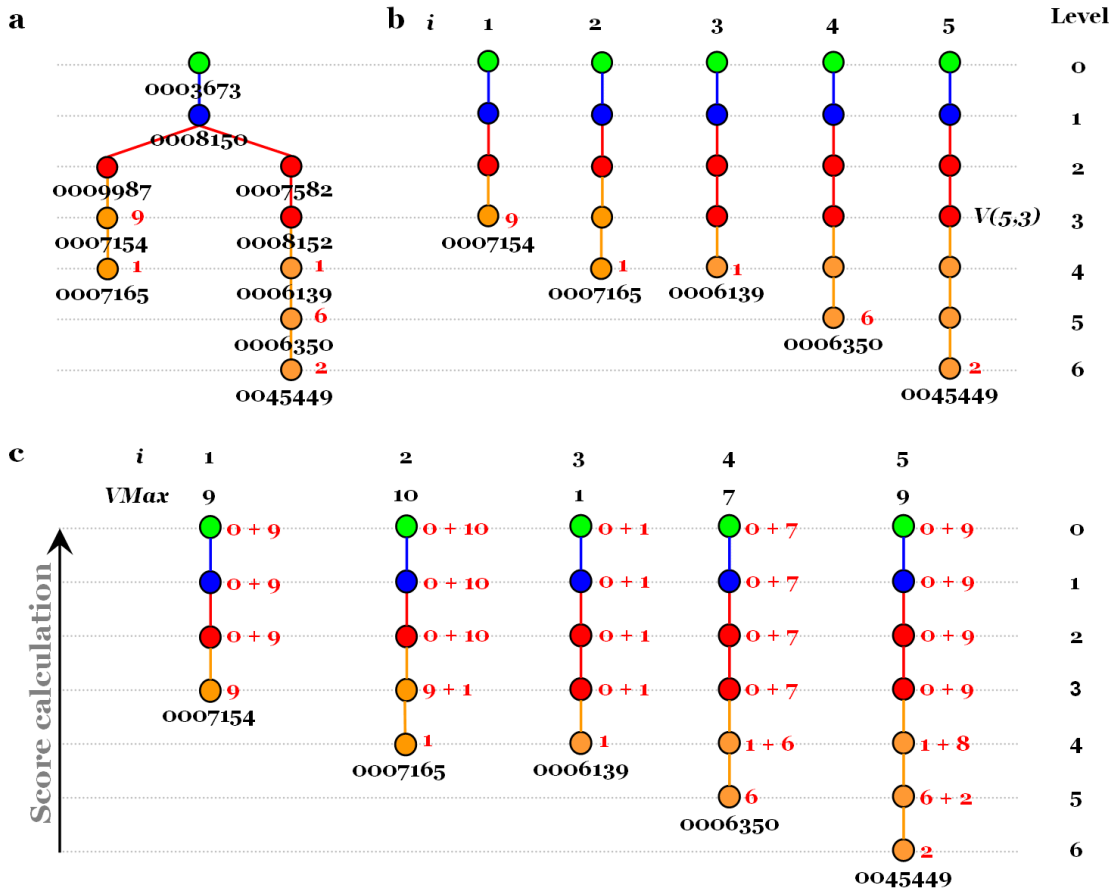


Figure 67. Determination of the scores allowing the construction of the MSO Mean Subfamily gene Ontology.
a – All the IPO terms (depicted in orange) allow the construction of the corresponding GO tree. The scores indicated in red are the number of proteins in the MACS assigned with this GO term as IPO. b – Decomposition of the branches from each IPO term to the root (depicted in green) of the tree. c – Determination of the V(i,j) score foreach node j of a branch I, of the VMax(i) score foreach branch I and of the VMAX score, maximal score among the VMax(i), VMax(2) in the present example.

A score is assigned to each IPO. It represents the number of proteins having this IPO. The scores are depicted in red on the Figure 67 a.

As depicted in Figure 67 b, the paths from each IPO to the root (GO level 0) are further decomposed forming as many linear branches as exists different IPO terms. For the j^{th} node of each i branch, a *V(i,j)* score is calculated from the highest GO level to the root of the tree. These *V(i,j)* scores represent the number of proteins annotated with this term as IPO, which is added to the score of the previous node *V(i,j+1)*. Considering a i branch having n(i) nodes, the maximal score of the branch i, *VMax(i)* and the maximal score of all the branches, *VMAX* are given by:

$$\mathrm{Max}_{i\in 1\ldots I}(\mathrm{VMax(i)}) = \mathrm{VMAX}$$

All these scores are detailed for each branch and each node in the example depicted in the Figure 67 c.

Several criteria then allow the selection of the GO terms which could reasonably be assigned to all the *ProtPop* proteins. First, a whole branch *i* is rejected if it does not contain enough proteins, this means less than *f* percent of *ProtPop*. Thus, if the following condition is verified:

$$VMAX * f > VMax(i)$$

This selection is further detailed for the example presented in Figure 68 a.

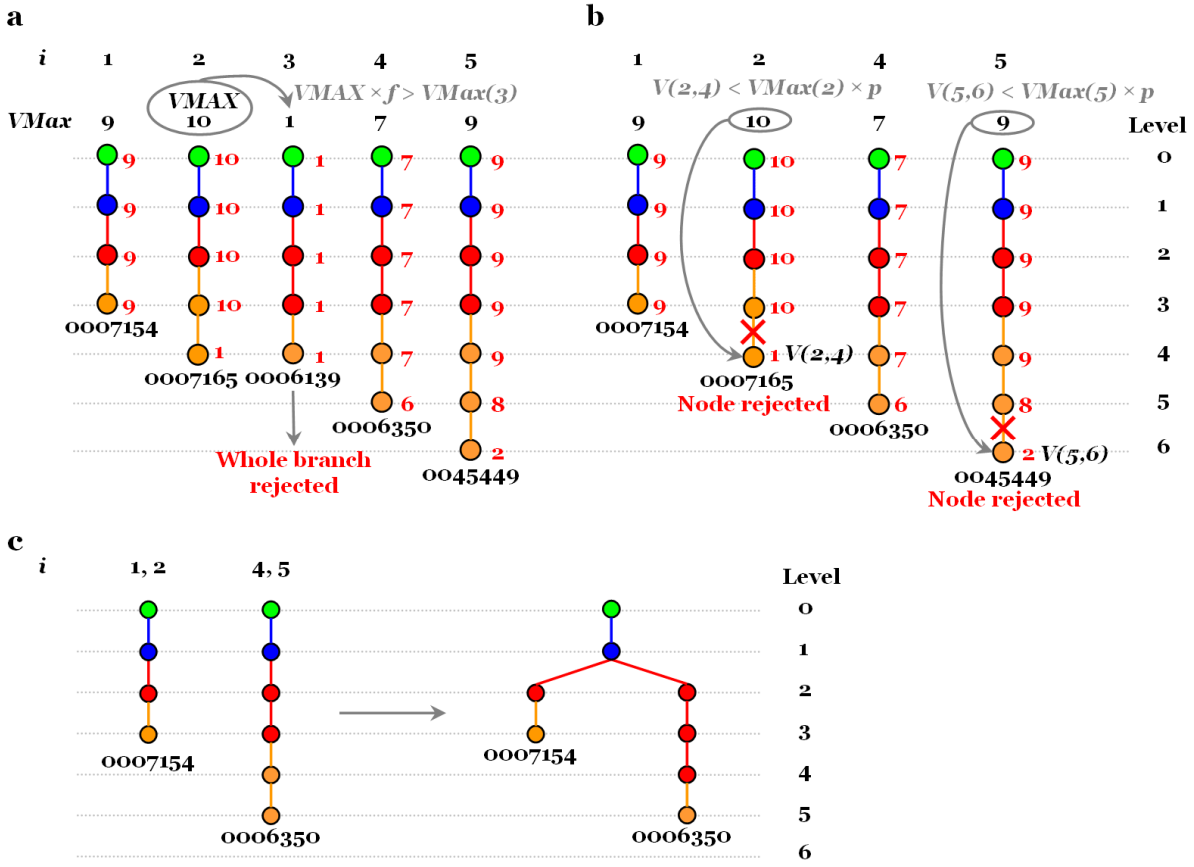


Figure 68. GO term selection with the GOAnno algorithm.
a – Complete removal of the branches containing not enough proteins compared to ProtPop. By default, f is set to 0,5. b – Rejection of the nodes that are too specialized. By default, p is set to 0,8. c - .Determination of the GO tree from the GO terms that have not been rejected during the two previous selection steps.

The second selection is the elimination of the nodes which are too specialized in a branch to be reasonably assigned to the set of the *ProtSubF* (Figure 68 b). Indeed, a node *j* of a branch *i* is rejected if it contains less proteins than p percent of all the proteins present in the branch i, as given by the following equation:

$$VMax(i) * p > V(i, j)$$

In the present example, both these selection steps eliminate the branch number 3, the node number 4 of the branch 2, and the node number 6 of the branch 5 of the GO tree. The GO terms which pass this selection define the MSO.

The fourth and last selection step is the determination of the GPO (Global Protein gene Ontology) defined by all the GO terms defining the IPO, PPO and MSO previously calculated (Figure 69), after removal of parent and redundant terms.

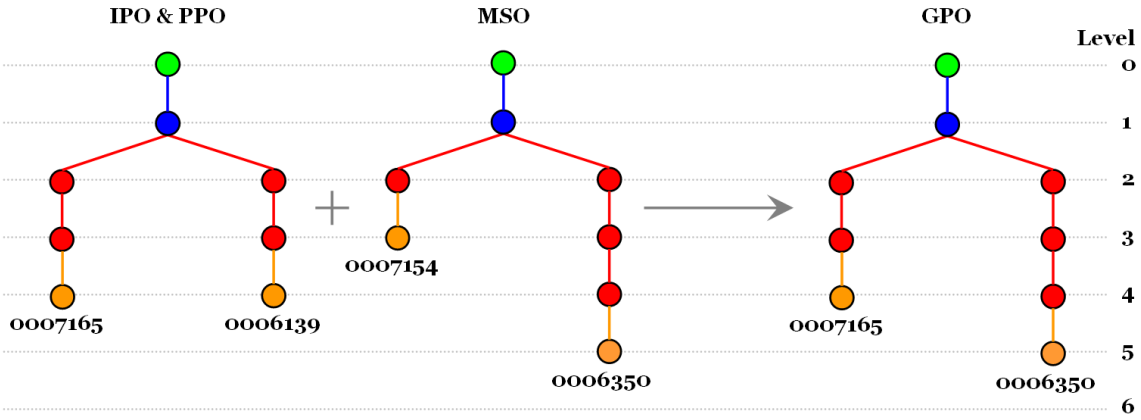


Figure 69. GPO, Global Protein gene Ontology definition. The GPO corresponds to all the GO terms defines by the IPO, PPO and MSO methods.

In the current example, the GPO finally corresponds to two GO terms depicted in orange and identified by the number 0007165 and 0006350, according to the Figure 69. We observe that, in comparison to the initial GO annotation (IPO), our treatment allowed us to improve the GO annotation of the studied protein by one GO level.

7.4 Availability and use

GOAnno is available as a local version for high-throughput analysis and is also available through a web server at http://bips.u-strasbg.fr/GOAnno.

Chapter 8 - Statistical enrichment estimation on group of co-expressed genes

Microarrays are at the center of a revolution in biotechnology, allowing researchers to simultaneously monitor the expression of tens of thousands of genes. Independent of the platform and the analysis methods used, the result of a microarray experiment is, in most cases, a list of genes found to be differentially expressed. The common challenge faced by researchers is to translate such lists of differentially regulated genes into a better understanding of the underlying biological phenomena. Based on the assumption that genes involved in biologically related processes frequently exhibit similar expression patterns, these genes should thus appear in the same cluster.

8.1 Approach

The common approach is thus to transform a list of differentially expressed genes into groups of co-expressed genes, according to their expression data, followed by the functional characterization of these groups with the Gene Ontology (GO) (Ashburner et al., 2000). This approach has been used to characterize groups of co-expressed genes in terms of biological process and involves highlighting biological processes, molecular functions or cellular components which are significantly enriched in the groups of co-expressed genes.

To obtain a broad description of the affected biological processes in a list of genes, one should take into account different functional annotations simultaneously, to describe as many aspects of the studied group as possible. The GO annotation is widely used for this task. Another approach is to characterize lists of co-expressed genes according to over-representation of specifc TFBSs in their corresponding promoter regions. This approach makes the assumption that groups of co-expressed genes may be co-regulated and may thus share the same regulatory pattern.

8.2 Implementation

The probability that a certain category occurs x times by chance in the list of differentially regulated genes is appropriately modeled by a hypergeometric distribution. However, the

hypergeometric distribution tends to the binomial distribution when the number of genes is large (Khatri and Draghici, 2005).

The problem in enrichment calculation can be described as follows: given a group of *R* entities (*R* group) among which r share a common chatacteristic and a global population of *N* entities (*N* population) among which n share this same characteristic Figure 70, is the considered characteristic overrepresented in the r entities of the group *R* compared to the n entities in the *N* population? In other words, is the ratio *r*/*R* significantly higher than *n*/*N*? To evaluate this comparison, we consider estimators based on both binomial (Doniger et al., 2003) and hypergeometric (Cho et al., 2001) statistical models.

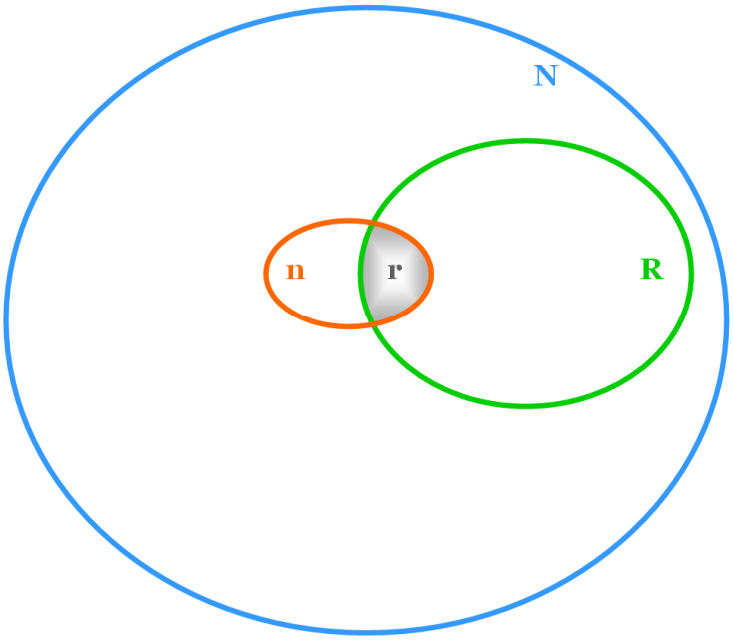


Figure 70. Representation of the entities implied in enrichment calculation.

To obtain the probability that a given term is enriched in a group of interest in comparison to a population, the first binomial statistical model calculates a *z score* that can be converted into a *p-value* using tables. The *z score* is calculated by subtracting the observed number of entities meeting a characteristic from the expected number of entities; and dividing by the standard deviation of the expected number of entities meeting the characteristic as given by (Doniger et al., 2003):

$$zscore = \frac{(\text{observed - expected})}{\text{Standard Deviation(expected)}} = \frac{\left(r - n\frac{R}{N}\right)}{\sqrt{n\left(\frac{R}{N}\right)\left(1-\left(\frac{R}{N}\right)\left(1-\frac{n-1}{N-1}\right)\right)}}$$

The choice of the global population *N* is of critical importance in the calculation of the significance of a potential enrichment. It is important to note that *N* can either be an entire population (as a proteome for example) or a random subset of an entire population.

Considering the populations *n*, *N*, *r* and *R* and defining X as the random variable equal to the number *k* ($k \leq n$) of entities having the considered characteristic, the hypergeometric probability law is then given by:

$$\mathrm{Prob}\{\mathrm{X} = \mathrm{k}\} = \frac{C_n^r \times C_{N-n}^{R-r}}{C_N^R}$$

The probability given by the hypergeometric law is a *p-value* assigned under the hypergeometric law. Nevertheless, the hypergeometric law becomes unusable when N is higher than several hundred. Thus, the hypergeometric distribution tends to the binomial distribution when the number of genes is large (Khatri and Draghici, 2005).

A positive value of *z score* reflects enrichment in the group *R* with comparison to the population *N*. Nevertheless, this enrichment is not always statistically significant. Indeed, in a context of *z score* calculation, a threshold above which the *z score* would reflect a statistical enrichment should be defined. Moreover, the n and r groups should contain enough entities to consider enrichment.

The obtained *z score* and *p-value* directly reflect the enrichment: high *z scores* or low *p-values* correspond to highly significant over-representations. A *p-value* lower than 0.05 is usually considered as highly significant.

8.3 Applications

8.3.1 Gene ontology

Enrichment calculations have been used to characterize groups of co-expressed genes in terms of biological process.

In this case, the whole human or mouse proteomes have been considered as a global population of comparison *N*; a considered group of co-expression constitutes the group *R* taken into account for the enrichment. The entities of these groups are gene products that have been annotated in the "biological process" ontology. In this context, the n entities can be the number of human proteins annotated in the GO term "*transcription regulation*". Thus, the *r* entities would be the number of proteins of the group of expression *R* that are

annotated with this GO term. An enrichment would defined as the proportion of entities *r* in the group *R* that is is higher than the proportion of entities *n* in the population *N*.

It should be noted that in such analysis, although genes can be represented by several probe sets in the microarray data, only a single copy of each gene or gene product has been taken into account in the GO analysis. Furthermore, the *z score* is calculated for all the GO terms assigned to each protein, including the parent terms.

8.3.2 TFBSs

To determine the relative overrepresentation of TFBS predictions between two sets of promoters, a control set and a set of interest have been calculated for several biological projects.

In the context of the TFBS analysis, two random data sets, each composed of 1000 mouse genomic sequences extracted from Entrez Gene IDs (see section 5.2.2) have been generated. They have been considered as population *N* because the TFBS predictions and the phylogenetic footprinting method have not been performed for the whole mouse Entrez Gene IDs because of time-calculation. The random set of sequences as a reference provides an alternative solution.

Results and discussion

Chapter 9 - PromAn results visualisation with PromAn GUI

In this chapter, we will present the display of the PromAn results using the PromAn GUI (Graphical User Interface) which allows a more in depth furter analysis of the results. This interface is also presented in the PromAn publication (see Annexe 1 -) but it is presented in more detail in the current section. The use of the PromAn GUI is described in the "Help" section (http://bips.u-strasbg.fr/PromAn/PromAnGUIHelp.html) on the PromAn web server.

As an example, we will use the Rhodopsin, the G-protein-coupled light receptor. This gene is expressed specifically in the rod photoreceptors of retina and plays an essential role in visual function. *Cis*-regulatory elements of the bovin rhodopsin proximal promoter region (RPPR) have been characterized in several studies (Chen et al., 1997) (Mitton et al., 2000).

Six vertebrate organisms; human, cow, dog, mouse, chicken and the frog *Xenopus tropicalis* were used to perform the PromAn analysis. For each organism, genomic sequences from – 10 kb to + 10 kb with respect to the start codon were extracted. The sequences were then aligned with the multiz program, using the mouse genomic sequence as the reference. The TFBS predictions were performed using 3 different databases: the public TRANSFAC database with the profile minimizing the false positive and false negative predictions (Match_TRANSFAC_MinSumGood), the JASPAR CORE database (Match_JASPAR_CORE) and the bibliographic database dedicated to the retina (Match_BIBLIO_Retina). The visualisation of the raw results with the PromAn GUI is illustrated in Figure 71.

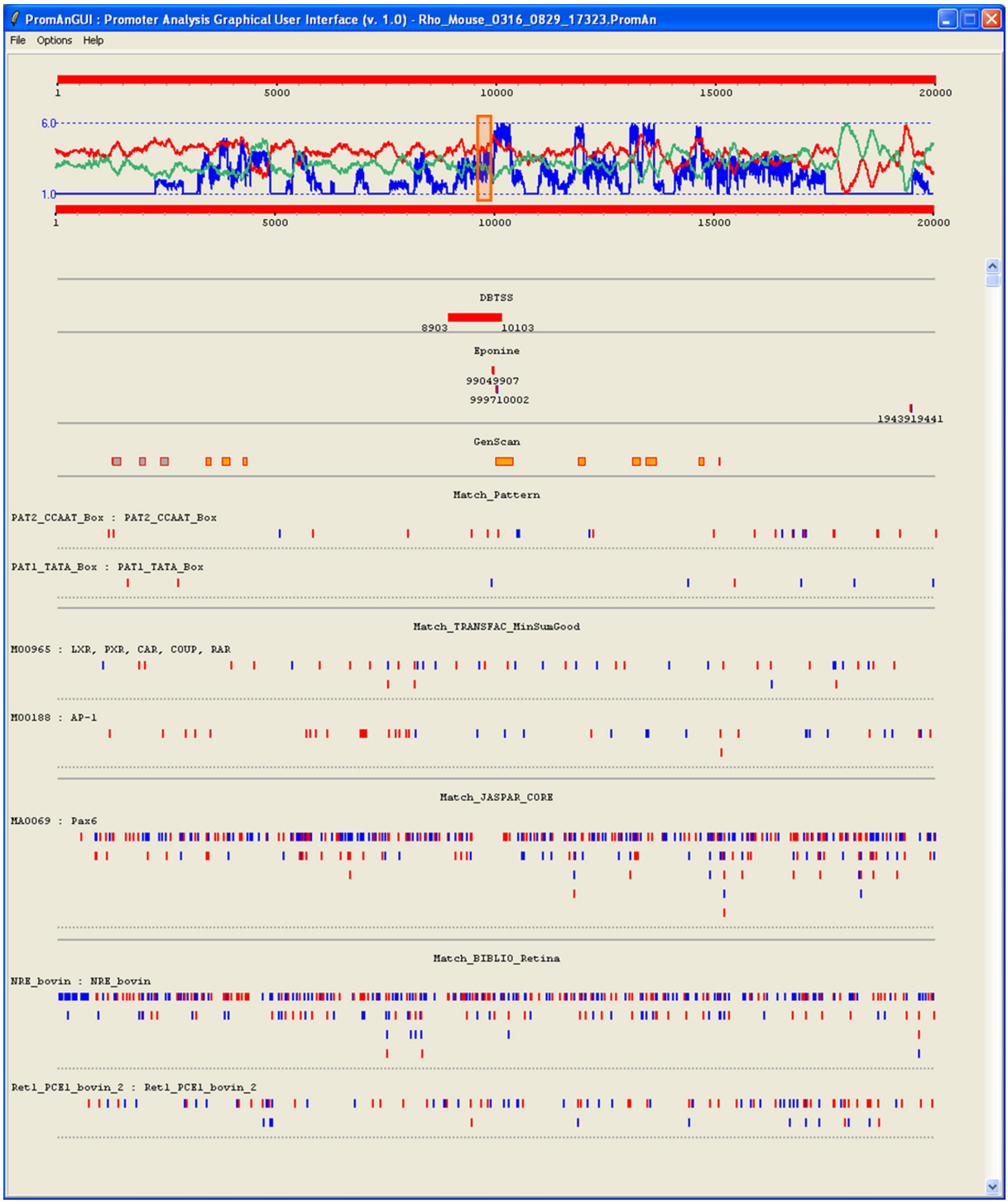


Figure 71. Visualisation of PromAn results of the rhodopsin analysis with the PromAn GUI. The upper frame displays the mouse rhodopsin genomic region (extracted from ±10,000 bp) with respect to the start codon with two red boxes indicating the scale of the visualized regions (here from 1 to 20,000 bp), surrounding two profiles: the dinucleotide profile (AT in green and GC in red) and the conservation profile (blue, scale from 1 to 6). The box in orange on the conservation profile highlights the RPPR described in Figure 72. The lower frames show, from top to bottom: the DBTSS, Eponine, GenScan, user-defined motifs (Match_Pattern), TRANSFAC (Match_TRANSFAC_MinSumGood), JASPAR (Match_JASPAR_CORE) and retina dedicated (Match_BIBLIO_Retina) predictions. The colors visualized here are the colors of the outline of the boxes (blue for minus and red for plus).

Results are always located with respect to the reference sequence (mouse), which is shown as two red boxes surrounding different profiles in the upper frame. Both AT (green) and GC (red) profiles are displayed to easily and immediately identify regions enriched in these dinucleotides. In this example, we can see that the mouse genomic sequence is globally GC rich while the rhodopsin gene is specifically expressed in eye and retina. Thus, the rhodopsin gene example clearly illustrates the observation that genes specifically expressed in brain are surprisingly enriched in CpG islands while most of the tissue-specific genes are AT rich (see section 3.2.3 . The blue curve represents the conservation profile of the reference sequence based on a multiple alignment of the orthologous genomic sequences. It highlights 5 peaks which are conserved in the six organisms considered. Typically, regions downstream of the start codon and conserved from mammals to frog correspond to coding exons. The conservation profile also shows a proximal region of about 1 kb long conserved upstream of the start codon in all four mammals. The DBTSS, Eponine, GenScan and TFBS predictions (user-defined TATA and CCAAT boxes, TRANSFAC with the MinSumGood Profile, JASPAR_CORE and BIBLIO_Retina databases) are shown in the lower frames by boxes colored according to the conservation to facilitate the visualization [gradient from low (grey) to high (red)] score. The DBTSS database allows the localization of an experimentally validated TSS at the 9,903th bp of the mouse reference sequence. The Eponine program also predicts a potential TSS close to this experimental TSS. The GenScan program confirms that the peaks conserved between the six species may correspond to exons. The TFBS predictions are also shown in the second frame. We observe that with no selection a huge number of TFBSs are predicted.

Within the proximal region of 1 kb conserved among the four mammals, a small region (from 9,677 to 9,916—depicted in orange on the conservation profile) has been described as being responsible for the photoreceptor specificity and is named the RPPR. The Figure 72 shows a zoom-in on the conserved RPPR region.

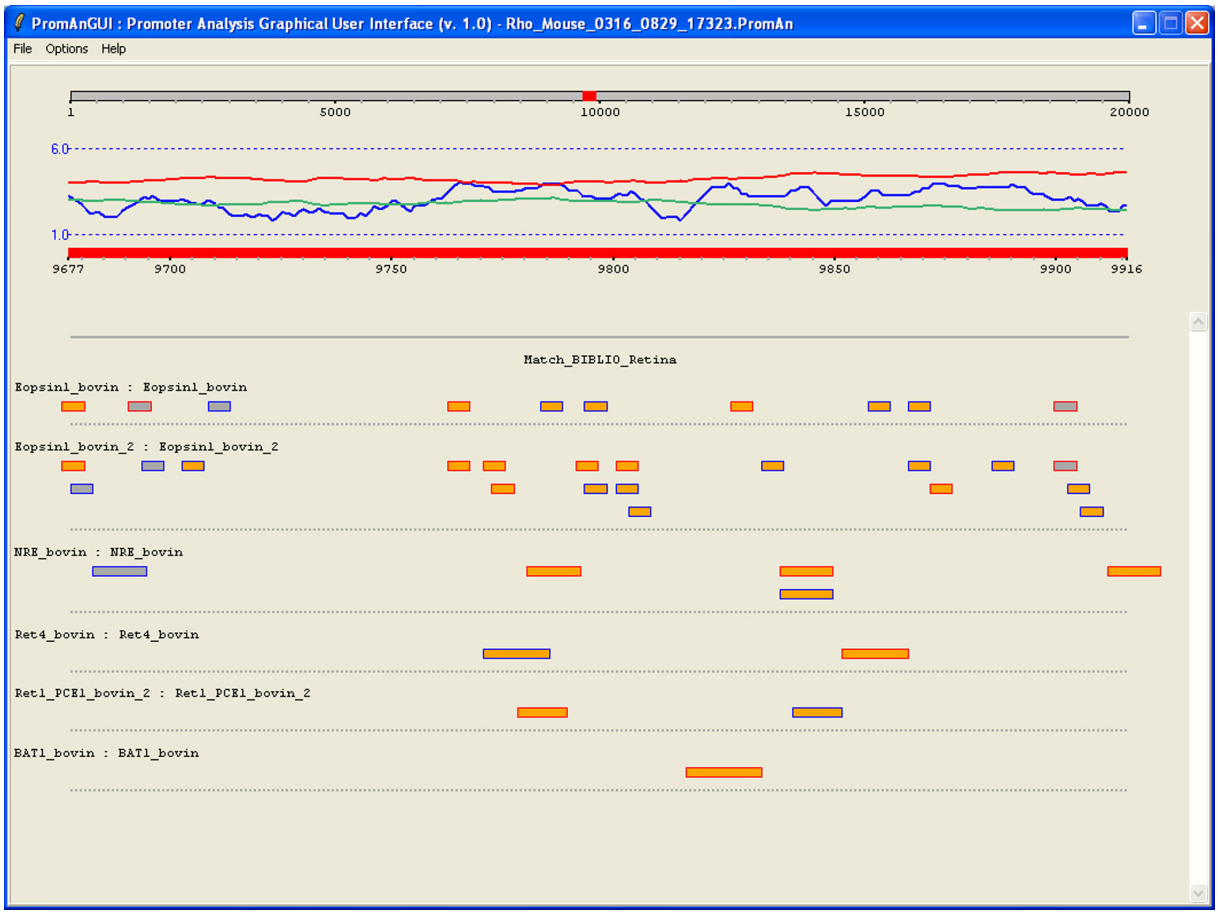


Figure 72. Analysis of the RPPR responsible for photoreceptor specificity.
The grey rectangle at the top displays the complete mouse reference sequence with the visualized region (zoom-in) colored in red. The red rectangle displays the scale of the region vizualized (from 9,677 to 9,916). The retina specific predictions are shown as a colored [gradient from low (grey) to high (red) conservation] box where the outline indicates the strand (blue for minus and red for plus). The blue dashed lines correspond to the scale of the conservation profile (see 6.1.2.3.2). It represents the number of organisms in which each nucleotide of the reference sequence is conserved.

The dinucleotide profiles show that the RPPR is a GC-rich region. The full genomic reference sequence is shown in grey with the zoomed-in region colored in red above the profiles. The scale of the zoomed-in region is displayed in red below the profiles. Matrix and conservation score cut-offs of 0.6 were used to select the TFBS predictions displayed. The Ret-1/PCE-1, BAT-1, NRE and Ret-4 *cis*-regulatory elements have been described in the bovin RPPR (Mitton et al., 2000) (Chen et al., 1997). This analysis shows that the RPPR is conserved during evolution. In the mouse region homologous to the bovin RPPR, PromAn retrieves the Ret-1/PCE-1, BAT-1, NRE and Ret-4 cis-regulatory elements. Thus, with this analysis, PromAn highlights the fact that these biologically active elements are conserved among mammals. We notice that many Eopsin-1 binding sites are predicted because the motifs are only 6 bp long.

Chapter 10 - Analysis of the general transcription factor TFIIH

10.1 Scientific context

The research group of Dr Jean-Marc Egly at the IGBMC is interested in the general transcription factor (TF) TFIIH, which is involved in the initiation of transcription, in particular by phosphorylating RNA polymerase II. TFIIH is also involved in the transactivation of hormonal nuclear receptors (NRs) (Keriel et al., 2002) (Compe et al., 2005) notably by inducing their phosphorylation, which is necessary for their activity. These NRs, through their binding to response elements present on many promoters, can recruit activators or co-activators, repressors or co-repressors as well as mediators which will regulate the expression of the downstream gene. This research group focuses their studies on the understanding of the transactivation mechanisms of the hormonal NR Peroxysome Proliferator-Activated Receptor (PPAR) by the general TF TFIIH.

PPARs regulate storage and catabolism of fats and carboxyhydrates. They are ligand-activated TFs and members of the nuclear hormone receptor superfamily. Three PPAR subtypes (α, β/δ, and γ) with a high degree of sequence conservation for each subtype across various species have been characterized. Within the superfamily of the NRs, PPARs belong to the subgroup which forms heterodimers with the 9-*cis*-retinoic acid receptor (RXR) and typically bind to DNA elements containing two copies of an idealized consensus binding site –AGGTCA– arranged in a Direct Repeat array spaced by n nucleotides (DRn) (Juge-Aubry et al., 1997).

To study the transactivation mechanism involving TFIIH TF and PPAR NR, Dr Pascal Trempat focuses on the fatty acid-binding protein-4 (FABP4) gene as a model while Dr Nicolas Le May centers his attention on the 3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (HMGCS2) gene model. The promoters of both genes have been analysed with the PromAn program and the results will be described in detail below.

10.2 Adipocyte Fatty Acid-Binding Protein 4, FABP4, gene

10.2.1 Scientific context

The adipocyte fatty acid binding protein (aP2) gene, namely fatty acid-binding protein-4 (FABP4) has been described as being regulated by the PPAR NR (Wahli et al., 1995). Through quantitative RT-PCR (Reverse Transcriptase Polymerase Chain Reaction) experiments in Hela cells, Pascal Trempat observed a transactivation of the endogenous aP2 gene with the PPAR NR after ligand addition. The aP2 gene was thus selected as an interesting model to study the transactivation mechanism involving at least PPAR and TFIIH TFs.

Molecular studies of the proximal promoter of the mouse FABP4 gene have defined two DNA elements which are important for the gene's expression, an AP1 and a C/EBP binding site (Christy et al., 1989) (Distel et al., 1987) (Herrera et al., 1989). By deletion analysis, Graves *et al.* identified a 520 bp enhancer specific to the adipose tissue at -5.4 kb of the mouse FABP4 gene (Graves et al., 1992). The Figure 73 depicts the structure of the FABP4 promoter.

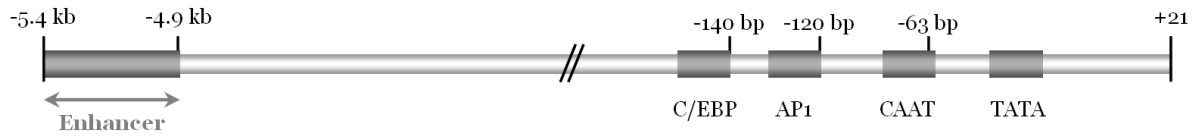


Figure 73. Structure of the mouse FABP4 promoter.
Segment from -5.4 kb to +21 of the FABP4 gene. The TSS is located at position 0. TFBSs for the AP1 and C/EBP TFs as well as the CCAAT and TATA box are indicated as boxes. In addition, the enhancer is shown from -5.4 kb to -4.9 kb. Adapted from Graves *et al.* (Graves et al., 1992).

PPARγ is predominantly expressed in adipose tissue, the small intestine, and lymphatic tissue. In coordination with C/EBP α and β, PPARγ is a key factor in inducing adipose differentiation (Tontonoz et al., 1995). The FABP4 gene is one late target gene of PPAR in the process of adipose differentiation from fibroblasts. The fat-specific enhancer present in the mouse FABP4 gene has been further described to bear two response elements, ARE6 and ARE7 for which PPARγ has a high binding affinity (Juge-Aubry et al., 1997).

The promoter of the FABP4 gene was described in mouse and in human in the 5,400 bp upstream of the TSS (Graves et al., 1992) (Rival et al., 2004).

The genomic region upstream of the TSS of the human FABP4 gene has been studied with PromAn. This analysis is presented in the following section.

10.2.2 PromAn analysis

The human promoter AJ627200 described in the literature (Rival et al., 2004) is located on the human contig extracted for PromAn analysis (Figure 74).

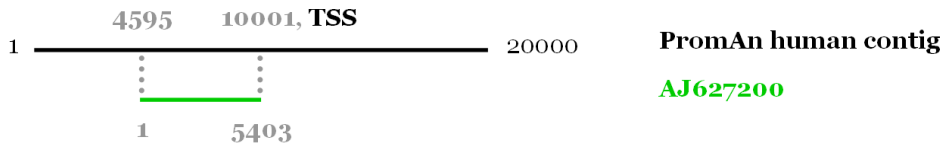


Figure 74. Location of the known human promoter AJ627200 on the PromAn human contig.

This promoter corresponds to the 5,403 kb upstream of the human TSS and is thus located from position 4,595 to position 10,001 on the PromAn human contig (20,000 bp in length, TSS at position 10,001).

The PromAn analysis was performed using five vertebrate organisms: human, chimp, rat, mouse and chicken. The conservation profile determined with PromAn is depicted in Figure 75 with the mouse and the human genomic sequences as reference.

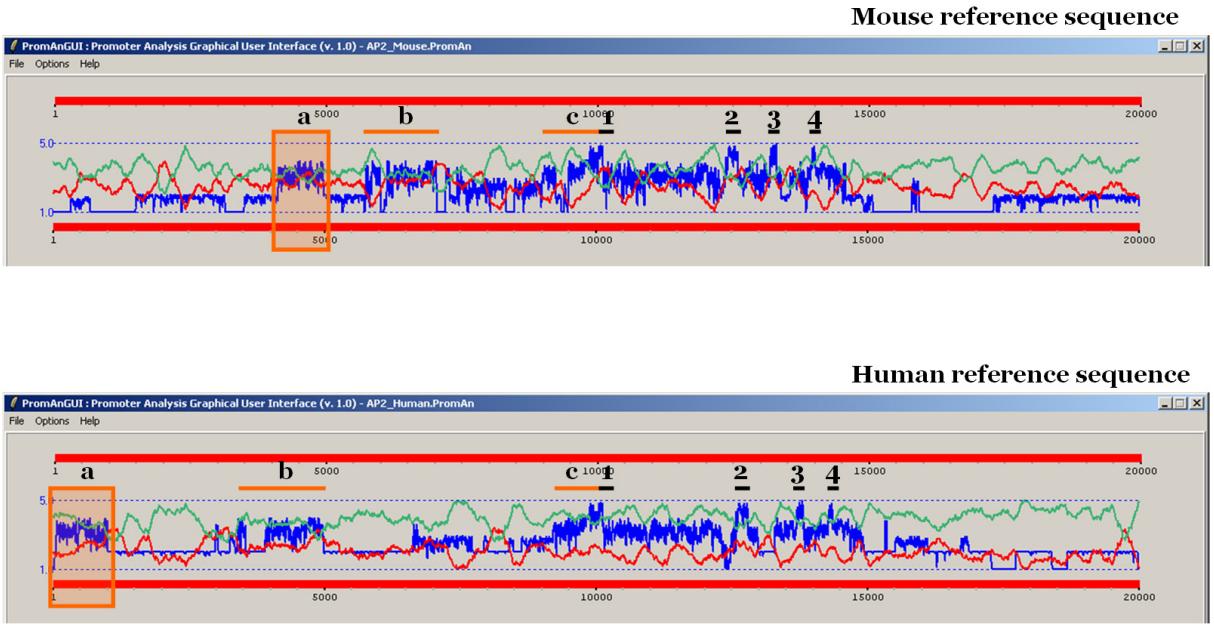


Figure 75. PromAn analysis of the FABP4 gene.
This analysis was performed using five vertebrates (human, chimp, rat, mouse and chicken), with the mouse and the human sequence as reference. In both cases, the TSS is located at position 10001 on the reference contig. The GC dinucleotide profile is colored in red and the AT dinucleotide profile in green and the conservation profile in blue. The numbered black lines localize the exons of the FABP4 gene as the peaks conserved in the five (scale indicated with the blue dashed lines) studied organisms. The regions of conservation upstream of the TSS are colored in orange and labelled from a to c.

The mouse and human FABP4 genomic sequences are globally AT rich with some small GC rich regions, such as the region around the TSS on the mouse sequence. This AT rich composition is in agreement with the general observation that most of the tissue-specific genes are AT rich (see section 3.2.3). The blue peaks conserved among the five organisms (scale indicated with the blue dashed lines) correspond to the exons numbered from 1 to 4 on the Figure 75. The proximal conserved region denoted "c" is highlighted in both conservation profiles. A detailed analysis of the multiple alignment showed that the conserved regions "a" and "b" on the mouse profile correspond to the conserved regions "a" and "b" on the human profile. For these conserved regions, the distances from the TSS are not conserved between mouse and human mainly due to the insertion of various repeats in the human genome. The region of conservation, outlined in orange, is located around -6 kb and -5 kb on the mouse reference sequence and was identified to be the mouse enhancer described in the literature (Graves et al., 1992). Furthermore, the detailed analysis of the multiple alignment of the region "a" highlighted the fact that the ARE6 and ARE7 elements described in the mouse are conserved in this human.

Thus, the analysis performed by PromAn predicts that the described mouse enhancer containing the elements ARE6 and ARE7 of the FABP4 gene is also present in the human FABP4 gene but at a position further from the TSS, namely at 9 to 10 kb upstream of the TSS. As a consequence, experiments were performed to verify the functionality of this potential human enhancer.

10.2.3 Experimental validation

To validate the location and activity of the human enhancer of the FABP4 gene, luciferase assays were performed.

Briefly, in these experiments, Hela cells were transfected with:

- The PPARγ2 NR cloned in the pSG5 vector.
- The pGL3 luciferase vector under control of the enhancer of the FABP4 promoter either from mouse or human.

The mouse enhancer considered here corresponds to the region containing the PPAR responsive elements ARE6 and ARE7 (mouse region "a" in Figure 75). The human enhancer tested in this experiment is the human genomic sequence identified as similar to the mouse enhancer during the PromAn analysis (human region "a" Figure 75). The murine elements, ARE6 (located from -5,320 to -5,290 related to the TSS) and ARE7 (located from -5,220 to -5,190 related to the TSS), have been described previously to be functional and to have a high

affinity for PPARγ (Juge-Aubry et al., 1997). However, the potential human enhancer located at -10 kb related to the TSS of the FABP4 gene has not been described in the literature. The results of this luciferase assay are shown in Figure 76.

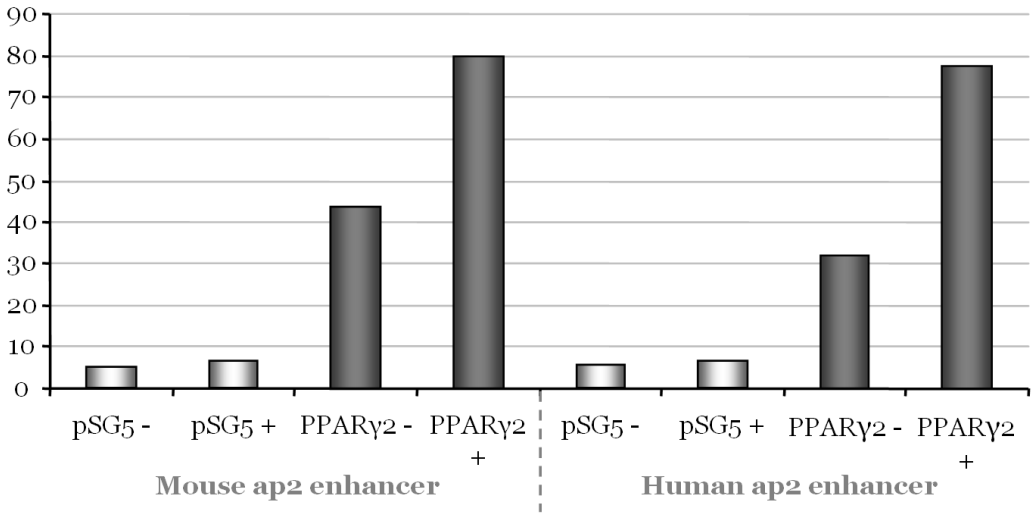


Figure 76. Luciferase assay with the mouse and human enhancers identified with PromAn.
The mouse and human ap2 enhancers correspond to the regions named "a" in Figure 75. "-" and "+" correspond to the absence or presence of the rosiglitazone PPAR ligand respectively. "pSG5" correspond to cells co-transfected with the pSG5 control vector and with the pGL3 luciferase vector under control of either the mouse or human FABP4 enhancer. "PPARγ2" correspond to the cells co-transfected with the pSG5 vector containing the PPARγ2 NR and with the pGL3 luciferase vector under control of either the mouse or human FABP4 enhancer. To estimate the level of transcription induced by the enhancer, the quantity of luciferase protein is measured.

In the presence of the PPAR ligand, the rosiglitazone, the pGL3 vectors which contain the mouse and human enhancers highly transcribe the luciferase gene. Therefore, the Hela cells treated with the rosiglitazone show that PPAR NR binds the ARE6 and ARE7 elements and thus allows luciferase gene transactivation. These experiments highlight the fact that the human enhancer identified with PromAn is functional.

Analogous luciferase assays have been performed with Hela cells transfected with a pGL3 vector containing only the 5.4 kb upstream of the TSS of the FABP4 human gene (AC018616) (Rival et al., 2004). This region, identified by analogy to the mouse FABP4 promoter, has been previously described as the potential human enhancer (Rival et al., 2004). The results are shown in Figure 77.

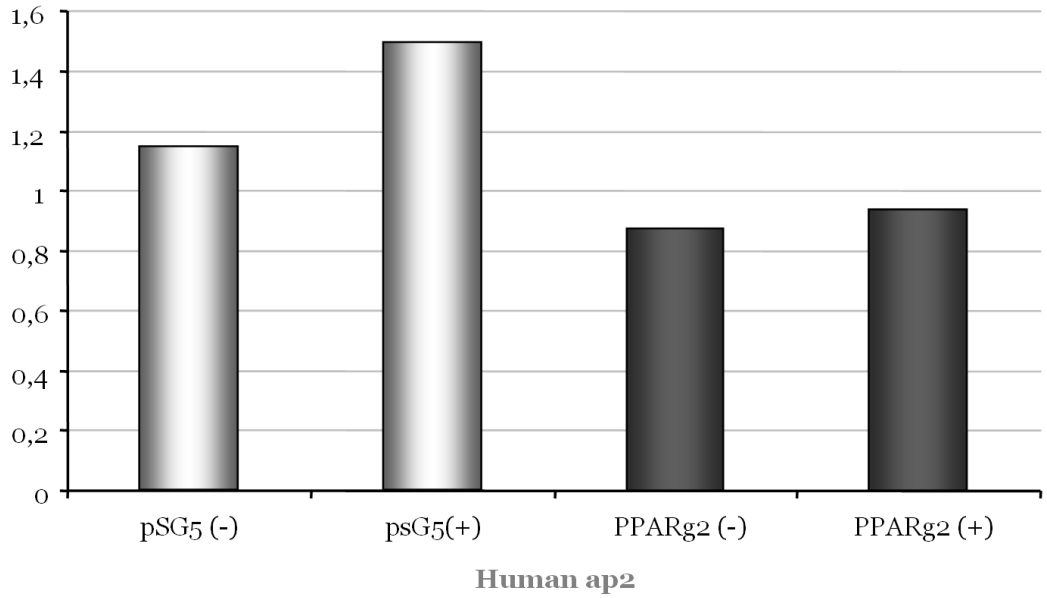


Figure 77. Luciferase assay with the human described FABP4 promoter.
See legend of the Figure 76. In this experiment, the vectors are under control of only the 5.4 kb upstream of the TSS of the FABP4 human gene.

In this experiment with the AC018616 vector, the level of luciferase observed in the presence of PPAR and the rosiglitazone ligand, is the same as that observed in cells non-treated with the rosiglitazone and is lower than the level observed in cells transfected with pSG5. This suggests that no PPAR responsive element is present in the 5.4 kb upstream of the human TSS. The transcription level observed here corresponds to the basal level of transcription, with no transactivation by PPAR.

In conclusion, the human sequence corresponding to the described mouse enhancer is not in the region of -5.4 kb upstream of the human TSS but is in fact the enhancer region identified with PromAn at -10 kb related to the TSS. Thus, the luciferase assays validated the PromAn results. This example demonstrates that the TSS of equivalent genes do not always reside at equivalent positions in the human and mouse genomes (Frith et al., 2006) and highlights the importance of a bioinformatics analysis with PromAn prior to experimental studies.

10.3 3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2, HMGCS2 gene

10.3.1 Scientific context

Some mutations in the TFIIH subunits are responsible for certain recessive genetic diseases (Trichothiodystrophy –TTD-, xeroderma pigmentosum –XP-, xeroderma pigmentosum and Cockayne syndrome combination -XP/CS-). In TTD, genes regulated by PPARα are deregulated. Hence, Dr Nicolas Le May is interested in understanding the role and mechanism of action of TFIIH in the transactivation of genes regulated by PPARα.

In the adult rat, PPARα is involved in the regulation of hepatic lipid metabolism and is mostly detected in liver, kidney, heart, brown adipose tissue and the intestine (Lee et al., 1995). To understand the role of TFIIH, the expression of the 3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (HMGCS2 or HMG-CoA synthase), a gene known to be regulated by PPARα is studied. This gene is expressed in liver and several extrahepatic tissues, such as the colon (Camarero et al., 2006). The proximal promoter of the rat HMGCS2 gene contains a functional PPAR response element named PPRE for for Peroxisome Proliferators Response Element (Rodriguez et al., 1994). This study showed by deletion analysis and mutation experiments that the PPRE is located between nucleotides -104 and -92 relative to the TSS. A functional PPRE was also identified in the proximal promoter of the human HMGCS2 gene by examination of the corresponding human region and was located from nucleotides -170 to -153 relative to the start site of translation (Hsu et al., 2001). The rat and human PPRE identified in the promoters of the HMGCS2 genes are presented in the Figure 78 (Hsu et al., 2001).

```
                  5′                              3′
ratHMGCS2          AAAAACT GGGCCA A AGGTCT CAG
humanHMGCS2        TAAAACT GGGTCA A AGGGCT CAC
```

Figure 78. PPRE sequences characterized in the rat and human HMGCS2 genes.

These elements were described to be targets of a PPARα/RXRα heterodimer (Hsu et al., 2001).

The HMGCS2 promoter was studied with PromAn and with footprinting experiments. The results of these analyses are presented in detail below.

10.3.2 PromAn analysis

The PromAn analysis was performed with the human HMGCS2 genomic sequence as reference. The analysis is based on five vertebrates: human, dog, cow, mouse and chicken. The contigs were extracted from -10 kb / + 10 kb with respect to the TSS (5′ end of the corresponding RefSeq mRNA). It is important to note that the CDS begins at position 10,052. The multiple alignment was constructed with the TBA program.

The Figure 79 shows the PromAn results visualized with the PromAn GUI.

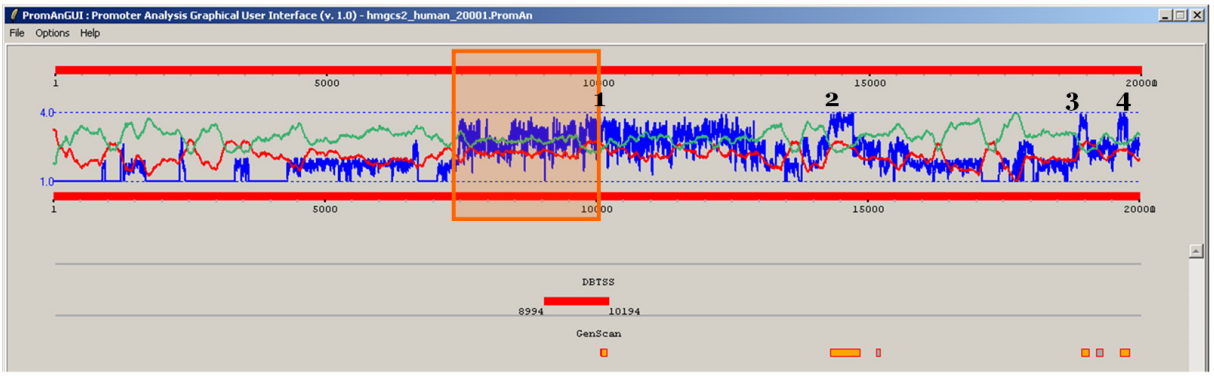


Figure 79. PromAn GUI for the HMGCS2 analysis with the human reference sequence.
The positions on the profiles are given relative to the human reference sequence. The human TSS is located at position 10001. The conservation profile (n) is shown in blue. According to the scale, the maximum of this profile is 4. Indeed, the chicken genomic sequence does not align with the human sequence. Above the profile, the black numbers indicate the location of the 4 first exons of the HMGCS2 gene. The GC dinucleotide profile is colored in red while the AT dinucleotide profile is in green. The DBTSS hit is also shown as well as the GenScan prediction results. The conserved region upstream the TSS is colored in orange.

The human genomic sequence is globally AT rich with local exceptions including the region flanking the TSS.

The DBTSS identifies a TSS at position 9,994, at 7 bp upstream of the potential TSS (position 10,001). Thus, the location of the TSS and the region to be studied are confirmed as potential promoter region. The GenScan program predicts exons with high scores in orange. These predictions correspond to the location of the known exons indicated with black numbers. It is important to note that GenScan does not predict a potential 5′ non coding exon. This constitutes a further confirmation of the location of the TSS and thus of the promoter region.

This conservation profile is very atypical, firstly because a large region around the TSS is so highly conserved that we do not distinguish the first exon and secondly, because immediately upstream of this region, the conservation is dramatically reduced. The regions located around the exons are highly conserved among the four organisms. Nevertheless the

region of conservation around the first exon is very large, from roughly position 7,500 to position 13000 on the human genomic sequence. This allows us to hypothesis that the region 2.5 kb upstream of the TSS (from position 7,500 to 10,000) as well as the region immediately downstream of the first exon to the position 13,000 constitute a potential active promoter.

We focus in more detail on the promoter region upstream of the TSS, depicted in orange in the Figure 79. The following figures (Figure 80 and Figure 81) therefore show a zoom-in on the region from 7,000 to 10,100. The conserved region upstream of the TSS is always highlighted in orange.

Figure 80. PPAR TFBS predictions for the conserved region (colored in orange) with no score selection.
TFBS predictions were performed on the Vertebrate Non Redundant profile of matrices available in the professional TRANSFAC database. Two matrices allow the prediction of PPREs: M00762 and M00763.

This zoom allows a higher resolution of the conservation in this region. Globally, we observe that the whole region seems to be well conserved with a lower conservation in the region ranging approximately from position 8,000 to 8,400.

Furthermore, we notice the huge amount of PPRE predictions obtained for a genomic sequence when no selection is applied. The following figure shows the PPRE predictions which passed a stringent selection step: matrix score and PromAn conservation score both higher than 0.8.

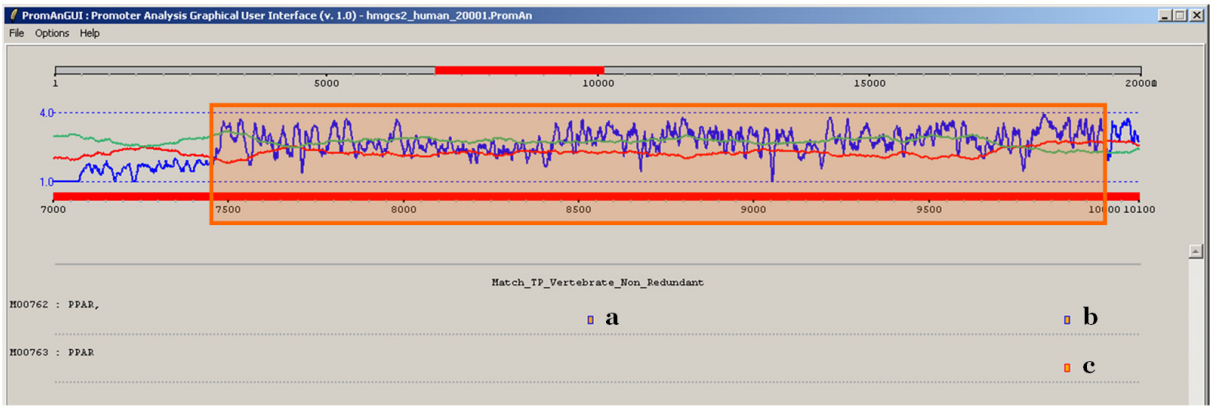


Figure 81. PPAR TFBS predictions for the conserved region (colored in orange) with stringent score selection.
In this region, predicted PPREs were selected with a matrix score (match motif score) and a conservation score both higher than 0.8.

Finally, we observe that only three PPAR TFBS predictions remain after this stringent selection. Two sites were predicted on the minus strand (blue borders, denoted **a** and **b**) and one site (red border, denoted **c**) was predicted on the plus strand. The characteristics of these PromAn predictions are shown in Table 20.

	A	**b**	**c**
ID	M00762	M00762	M00763
Name	PPAR	PPAR	PPAR
Strand	-	-	+
Start	8521	9882	9882
End	8533	9894	9894
Match core score	0.904	1.000	0.888
Match motif score	0.847	0.897	0.913
Conservation score	0.827	0.808	0.808

Table 20. PromAn information for the predicted PPAR TFBSs.

We observe that the motifs **b** and **c** correspond to the same TFBS located at the same position (from 9,882 to 9,894). This site corresponds to the PPRE previously described in the literature. The predicted motifs are 13 bp long while the described motifs are 23 bp long. We can validate the nucleotide sequence of these predicted motifs in PromAn (Figure 82).

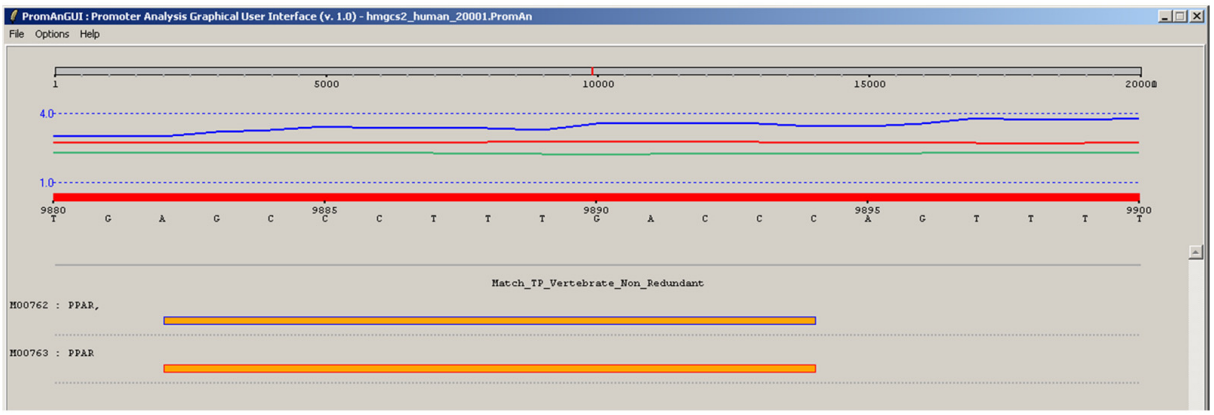


Figure 82. PromAn zoom-in on the predicted PPARs, which are proximal to the TSS.

The multiple alignment provided by TBA shows the blocks of conservation present in this proximal region of the promoter (Figure 83) including a highly conserved TATA-box and the PPAR TFBS.

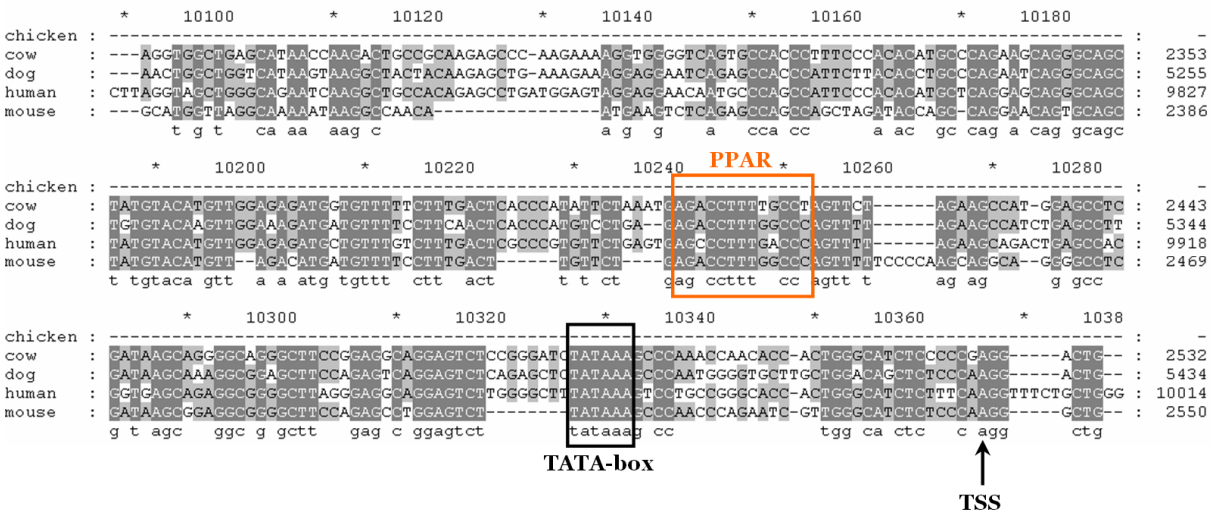


Figure 83. Multiple alignment of the proximal promoter of the HMGCS2 gene.
This multiple alignment was constructed with TBA from the human, mouse, cow, dog and chicken genomic sequences, with the human sequence as reference.

The motif **a** identified through the stringent selection step among a huge number of PPAR TFBS predictions, constitutes a major candidate for further experimental validation. No such PPRE in the region upstream -1.5 kb of the TSS is described in the literature. We hypothesise that this PPRE could represent an enhancer region which would be brought into the vicinity of the proximal promoter region through a "DNA looping mechanism" (see Figure 24). Such a model was described for the androgen nuclear receptor (AR) (Wang et al., 2005) which regulates the prostate specific antigen (PSA) gene through binding to a proximal promoter region coupled to the binding of an enhancer region approximately 4 kb upstream. Recruitment of AR and the coactivators at both sites was described as creating a

chromosomal loop that allows RNA polymerase II to track from the enhancer to the proximal promoter. Experiments are currently in progress to investigate the presence of an enhancer region in the promoter of the HMGCS2 gene.

Finally, the highly conserved region immediately upstream of the TSS (motif b/c) was further studied by DNase I footprinting experiments, the results are presented in the next section.

10.3.3 DNase I footprinting experiments

DNase I footprinting is a commonly-used technique for the identification of the specific DNA sequence to which the TF binds. This method takes advantage of the fact that when a protein is bound to a region of DNA, it protects the DNA sequence from digestion by nucleases.

Briefly, labeled samples of a DNA fragment are digested in the presence or absence of a DNA-binding protein, denatured, electrophoresed and the resulting gel is subjected to autoradiography. The region protected by the bound protein appears as a gap, or "footprint," in the array of bands resulting from digestion in the absence of protein. When footprinting is performed with a DNA fragment containing a known DNA control element, the appearance of a "footprint" indicates the presence of a transcription factor that binds the control element in the protein sample assayed. This method is briefly described in Figure 84.

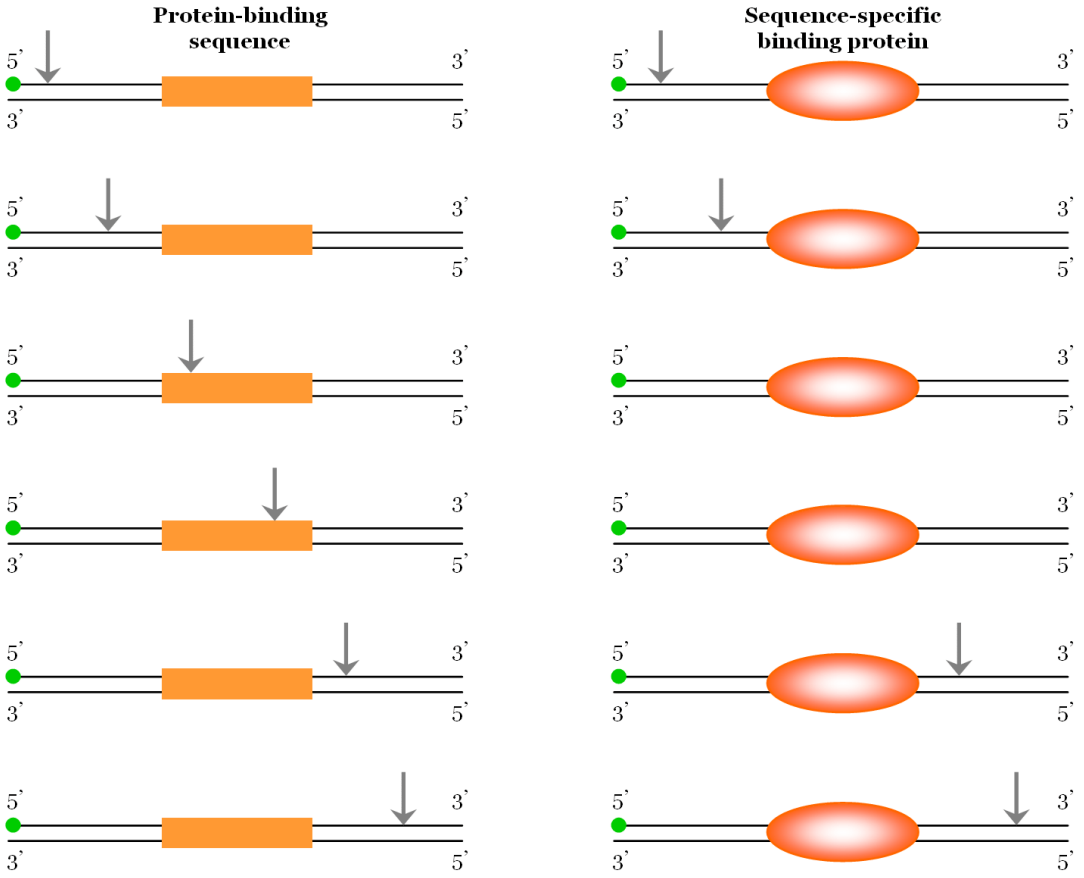


Figure 84. Description of the DNase I digestion in the DNase I footprinting method. Adapted from "Molecular cell biology", Lodish *et al.* fifth edition, Figure 11-13.

The DNA fragment from -260 to + 60 relative to the human HMGCS2 TSS was labeled at one end with ^{32}P (green dot). Portions of the labelled DNA sample are then digested with DNase I in the presence and absence of either purified PPAR/RXR heterodimers or nuclear extracts from Hela cells supplemented with purified PPAR/RXR heterodimers. Indeed, RXR is endogeneously expressed in Hela cells, but PPAR is not expressed. Nevertheless, both PPAR and RXR are supplemented in Hela cells in order to respect a given stoechiometry between PPAR and RXR. The resulting gels are shown in Figure 85.

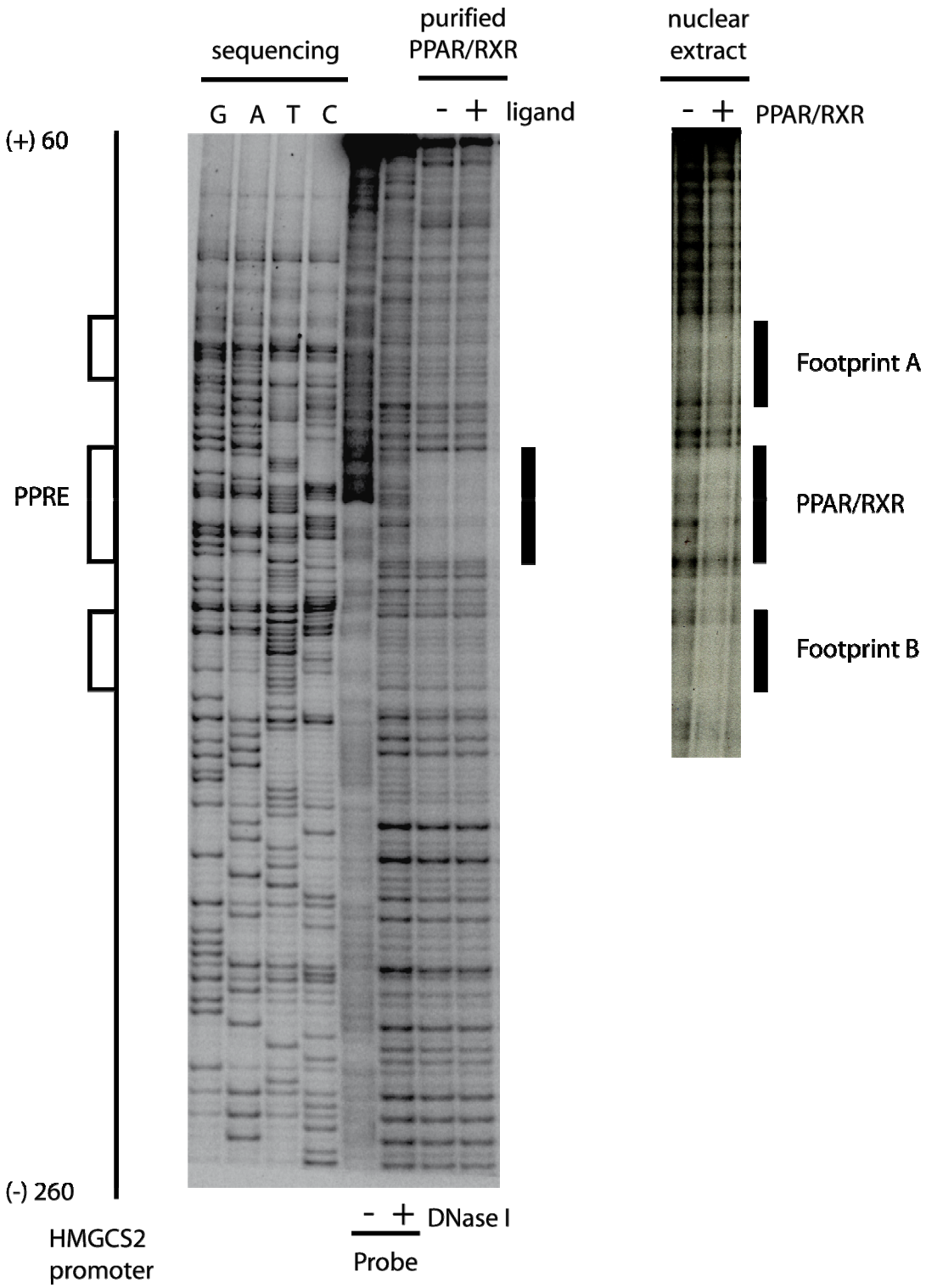


Figure 85. DNase I footprinting of the human HMGCS2 promoter.
The DNA probe used for this footprinting analysis is the human genomic sequence of the HMGCS2 promoter from -260 to +60 relative to the TSS. "Probe DNase I -" corresponds to a control with the DNA probe alone, without DNAse (DNA probe do not migrate far into the gel and bands correspond to random degradation of the probe). The "Probe DNase I +" corresponds to the DNase I digestion in absence of the DNA binding protein. "Purified PPAR/RXR" corresponds to the DNA probe in presence of purified PPAR/RXR heterodimer with or without ligand ("+" and "–", respectively). Finally, right panel, "nuclear extract PPAR/RXR -" corresponds to the DNase I experiments in presence of nuclear extract without (-) or with (+) supplementation with purified PPAR/RXR heterodimer.

Cleavage bands are missing in the presence of purified PPAR/RXR. These missing bands on the gel constitute the "footprint" identifying the PPAR, RXR, or PPAR/RXR heterodimer binding region. A similar footprint region is highlighted in the presence of nuclear extract of Hela cells supplemented with PPAR/RXR heterodimer. Thus, this footprint confirms the protecting area of the PPAR Response Element (PPRE) described in the literature and predicted by PromAn. Further experiments validate the fact that the PPRE footprint is not identified in the presence of either PPAR or RXR alone (data not shown), i.e. that the heterodimer formation is necessary. Thus, we conclude that this human PPRE element is the target of a PPAR/RXR heterodimer.

According to the probe sequencing, the PPAR/RXR binding was precisely localised from 9,877 to 9,901. Further experiments are currently in progress to validate this location. Other regions are also protected in presence of nuclear abstracts independently of the presence of PPAR/RXR (see Figure 85). The footprint region "Footprint A" upstream of the PPAR element is located from -79 to -56 (position 9,922 to 9,945). Furthermore, a third footprint region "Footprint B" is located downstream of the PPRE from -149 to -135 (position 9,852 to 9,866). These footprint regions are shown on the genomic sequence multiple alignment in the Figure 86.

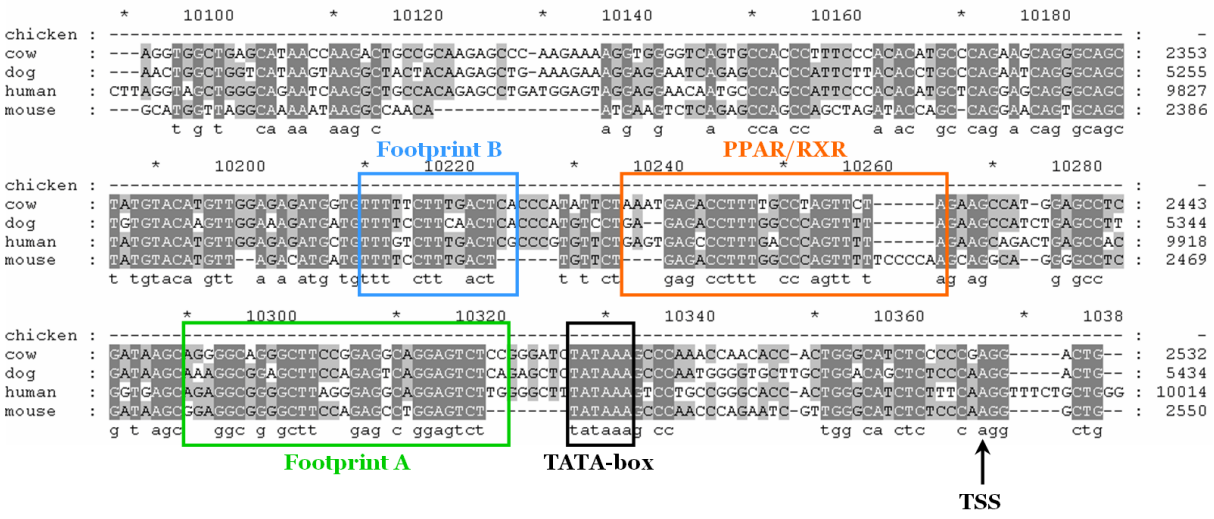


Figure 86. Localisation of the TFBSs identified from the footprint analysis.

Supplementary footprint experiments are now in progress to validate the exact locations of these footprint regions. Nevertheless, we have investigated the TFBSs predicted with PromAn in these genomic regions.

10.3.4 PromAn analysis of the "footprint" regions

10.3.4.1 Footprint A located from positions 9,922 to 9,945

The PromAn TFBS predictions are presented in the Table 21.

TP matrix	TP name	Official symbol	Gene ID	Start	End	Sense	Matrix score	Cons. score	Expression in liver
M00444	**VDR**	VDR	7421	**9919**	**9933**	+	0.765	0.800	**Yes**
				9932	**9946**	+	0.747	0.767	
				9933	9947	+	0.700	0.767	
				9934	9948	+	0.763	0.767	
M00800	AP-2	TFAP2A	7020	**9916**	**9931**	+	0.797	0.750	No
				9929	**9944**	-	0.789	0.812	
M00428	E2F-1	E2F1	1869	9926	9933	+	0.721	0.875	**Yes**
M00255	GC box			9925	9938	+	0.930	0.911	
				9937	9950	+	0.795	0.768	
M00446	Spz1	Spz1	84654	**9916**	**9930**	+	0.726	0.733	No
				9926	**9940**	+	0.892	0.783	
				9928	**9942**	+	0.763	0.783	
				9929	**9943**	+	0.803	0.800	
				9937	**9951**	+	0.738	0.783	
				9938	9952	+	0.879	0.783	
M00695	ETF	TEAD2	8463	9927	9933	+	0.996	0.893	No
M00008	**Sp1**	SP1	6667	**9927**	**9936**	+	0.956	0.900	**Yes**
				9932	**9941**	+	0.805	0.700	
				9942	**9951**	+	0.880	0.925	
M00986	Churchill	CHURC1	91612	9931	9936	+	0.993	0.917	**Yes**
M00315	GEN_INI			9922	9929	+	0.745	0.781	
M00684	XPF-1			9919	9928	-	0.879	0.725	
M00706	TFII-I			9941	9949	+	0.945	0.861	
M00983	MAF	MAF	4094	9938	9948	+	0.838	0.705	No
M00720	CAC-binding			**9927**	**9935**	+	0.911	0.889	
				9942	**9950**	+	0.866	0.917	
M00761	p53	TP53	7157	9943	9952	+	0.880	0.925	
M00965	LXR	NR1H		9911	9927	+	0.726	0.735	
				9917	9933	-	0.715	0.750	
M00646	LF-A1			9933	9940	+	0.809	0.719	
M00148	SRY			9932	9938	-	0.744	0.929	
M00484	Ncx	TLX2	3196	9933	9942	-	0.732	0.725	**Yes**
M00991	CDX	TLX1	3195	9936	9953	-	0.712	0.819	No
M00623	Crx	CRX	1406	9931	9943	-	0.948	0.769	**Yes**
M00491	MAZR	PATZ1	23598	9926	9938	+	0.899	0.904	

Table 21. TFBSs predicted in the "Footprint A" region.
"TP matrix" is the access name of the matrix in the TRANSFAC Professional database. "TP name" is the name of the corresponding TF given in TRANSFAC Professional. "Official symbol" is the official symbol of the gene name. The official symbols were determined from Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=gene). Indication of TFBS localisation, sense and scores are given by PromAn. Matrix score and conservation score (Cons. score) higher than 0.7 were used to select the potential TFBSs. The liver expression (Liver expr.) were extracted from SymAtlas (http://symatlas.gnf.org/SymAtlas/) and from GermOnline (http://www.germonline.org/index.html) web servers. The TP names are not the "official" gene symbols. Thus for some of the genes no Gene ID could be identified and thus no liver expression information is given. Start and end positions are shown in bold when one or various TFBSs define a region encompassing the entire "Footprint A" region. The genes expressed in liver have a bold "Yes" in the last column.

Among the TFBSs predicted on the "Footprint A" region, we focus on the TFs that are able to bind the whole region (from position 9,922 to position 9,945) protected with proteins of the nuclear extract in the footprint experiment. The corresponding positions are shown in Table 21. Furthermore, TFs not expressed in liver are not good candidates because as previously described in section 10.3.1 , the HMGCS2 regulated gene is expressed in liver. Expression in liver is thus shown in bold in Table 21.

Finally, the location of the footprint region, the TFBS predictions and the TF gene expression in liver are combined to determine on the potential TFs that could bind to and protect the "Footprint A" region observed in the experiment. TFs validating both position and expression conditions are shown in bold in orange. This leads to the conclusion that the TF Sp1 and the nuclear receptor VDR (vitamin D receptor) are the best candidates for binding to the "Footprint A" region.

10.3.4.2 Footprint B located from position 9852 to position 9866

The TFBS predictions highlighted in PromAn are presented in the Table 22.

TP matrix	TP name	Official symbol	Gene ID	Start	End	Sense	Matrix score	Cons. score	Expression in liver
M00428	E2F-1	E2F1	1869	9862	9869	+	0.798	0.875	**Yes**
M00396	En-1	EN1	2019	9855	9861	+	0.891	0.714	No
M00806	**NF-1**			**9851**	**9867**	-	0.744	0.824	**Yes**
M00255	GC			9859	9872	-	0.803	0.804	
M00446	Spz1	Spz1	84654	9857	9871	-	0.768	0.833	No
M00008	Sp1	SP1	6667	9861	9870	-	0.842	0.775	**Yes**
M00716	ZF5	ZFP161	7541	9862	9869	-	0.739	0.875	**Yes**
M00033	p300			9844	9857	+	0.768	0.804	
M00315	GEN_INI			**9851**	**9858**	-	0.767	0.812	
				9857	**9864**	-	0.778	0.938	
M00983	MAF	MAF	4094	9859	9869	-	0.903	0.864	No
M00761	p53	TP53	7157	9860	9869	-	0.863	0.850	
M00148	SRY			9854	9860	-	0.868	0.750	
M00468	AP-2rep	KLF12	11278	9849	9855	+	0.783	0.714	No
M00192	**GR**	NR3C1	2908	**9843**	**9861**	+	0.864	0.816	**Yes**
				9860	**9878**	+	0.928	0.825	
M00791	**HNF-3**	FOXA		9848	9860	-	0.964	0.788	**Yes**
				9854	**9866**	-	0.855	0.769	
M00790	**HNF-1**	TCF1	6927	**9849**	**9866**	+	0.785	0.791	**Yes**
M00456	FAC1	FALZ	2186	9846	9859	-	0.747	0.821	**Yes**
				9847	9860	-	0.774	0.804	
M00445	Xvent-1			9846	9858	+	0.833	0.808	
				9852	**9864**	+	0.736	0.846	
M00183	c-Myb	MYB	4602	9847	9856	-	0.927	0.750	No
M00134	**HNF-4**	HNF4		**9850**	**9868**	-	0.710	0.842	**Yes**
M00978	LEF1TCF1			**9856**	**9866**	+	0.876	0.818	

Table 22. TFBSs predicted in the "Footprint B" region.
See Table 21 for legend.

Similarly to the analysis of the "Footprint A" region, predicted TFBS positions and expression in liver of the corresponding TF are considered in order to further analyse the TFBS predictions. GR TF is indicated in bold but not in orange because it is predicted to cover a DNA binding region 10 bp larger than the "Footprint B" region. As in the study of the "Footprint A" region, the TFs indicated in orange are good candidates for binding to the region "Footprint B". However, as described in section 10.3.1 the HMGCS2 gene is expressed in large colon and highly expressed in liver. We thus searched for transcription factors known to present a similar regulation expression profile. HNF-4 presents a similar expression profile with high expression in large intestine, liver and kidney, suggesting that HNF-4 constitutes the best candidate for binding in the "Footprint B" region.

A bibliography search highlighted the fact that Sp1 and HNF-4 were described as directly interacting in order to transactivate the apolipoprotein CIII promoter (Kardassis et al., 2002). Thus, this known physical interaction between Sp1 and HNF-4 strongly supports the hypothesis that these two TFs represent the best candidates to explain the experimental "Footprint A" and "Footprint B" observed in the nuclear extract. In the future, experiments will be performed to test this hypothesis.

Taken together, these two studies constitute good examples of the efficiency of the PromAn program and of the different types of analysis that can be performed with PromAn. First, they highlight the importance of the study of the conservation profile during evolution to determine the genomic regions that may correspond to potentially active promoter. This allows us to focus the experimental validations on these precise regions. Secondly, the study of the HMGCS2 gene illustrates our approach to the analysis of the PromAn TFBS predictions. The conservation and matrix scores are of crucial importance in the selection of the TFBS predictions. It is also interesting to include other information such as expression and interactomics data to propose a hypothesis of a potential regulatory network involved in gene expression regulation. As a final conclusion, these two studies highlight the advantages of an integrative approach for promoter analysis such as the approach developed in the PromAn program.

Chapter 11 - Analysis of *retinitis pigmentosa*

11.1 Scientific context

Retinitis pigmentosa (RP) is a genetically heterogeneous retinal degeneration characterized by the sequential degeneration of rod and cone photoreceptors. The first clinical signs of RP are night blindness and narrowing of the peripheral field of vision which progressively worsens to become "tunnel-like". Eventually, the central vision is reduced to complete blindness in most cases. Briefly, the retina is composed of six neuronal cell types spatially subdivided into laminated layers. Photoreceptors, which constitute 70% of retinal neurons, comprise 97% of rods and 3% of cones in mouse retina. At a cellular level in the RP, the retinal rod photoreceptors involved in night and side visions slowly degenerate. Subsequently, the cone photoreceptors responsible for both color and high-contrast visions, visual acuity, detail perception and normal light vision are similarly affected. To date, no treatment is available.

As cones are responsible for the most crucial visual functions, the mechanisms that trigger their degeneration are major therapeutic targets. The retinal degeneration 1 (*rd1*) mouse is the most studied animal model for the RP human disease. It carries a recessive mutation in the rod-specific cGMP phosphodiesterase beta subunit gene leading to rod photoreceptor apoptotic death (Chang et al., 1993) (Portera-Cailliau et al., 1994) followed by cone death presumably through lack of trophic support (Mohand-Said et al., 1998). The laboratory of Dr Thierry Léveillard works on the RP disease. Expression cloning was used to identify a trophic factor secreted by rods that promotes cone viability in the *rd1* mouse; RdCVF, for Rod-derived Cone Viability Factor (Leveillard et al., 2004). In the model proposed, rod degeneration results in the decrease of RdCVF expression and finally to cone degeneration by lack of trophic support. Genomic investigations of RdCVF revealed the presence of a paralogous gene, RdCVF2 (Chalmel *et al., manuscript in preparation*). RdCVF and RdCVF2 present similar gene structure. Furthermore, the cone trophic factor activity of RdCVF2 was found to be similar to that of RdCVF.

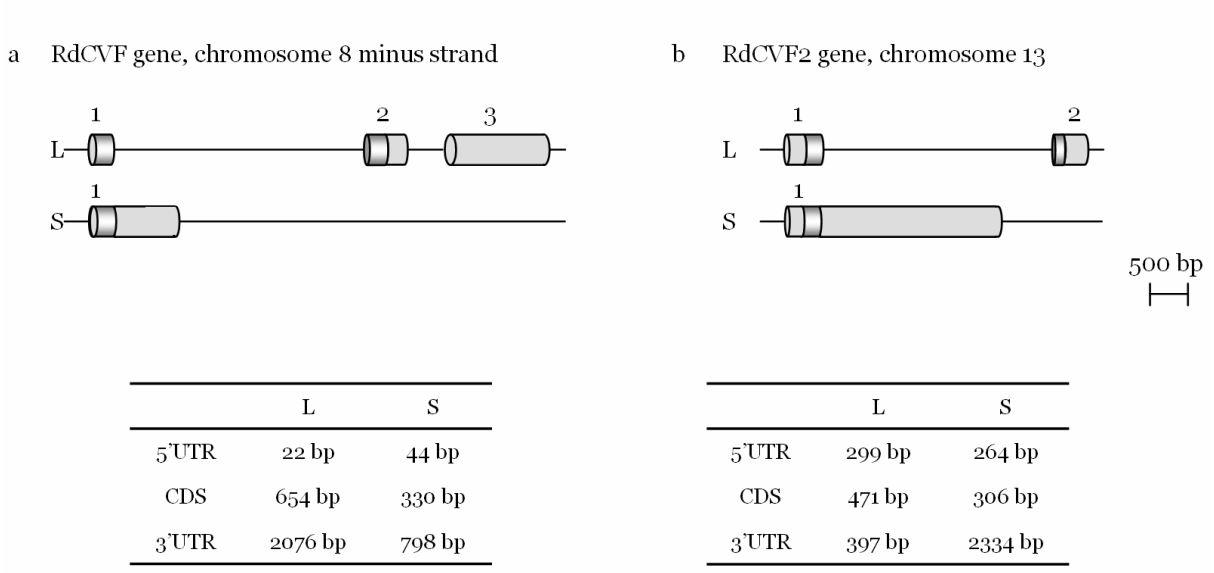
Since cone survival is critical to maintain useful vision for patients with retinal degeneration, RdCVF and RdCVF2 may provide keys to develop treatment for the RP disease. So, the identification of the regulatory mechanisms controlling RdCVF genes expression is crucial

for the understanding of the function of these genes and for the development of potential therapeutic agents. The PromAn program has been used for the analysis of the RdCVF gene promoters.

After a brief description of the RdCVF gene structure, their promoter analysis with PromAn will be presented, followed by experimental validations performed by Dr Sophie Lambard in the laboratory of Thierry Léveillard.

11.2 Gene structure

The RdCVF gene is transcribed into two mRNA, a long RdCVF-L and a short RdCVF-S form. These transcripts encode for a long protein form having a putative thiol-oxydoreductase activity (Jeffery, 1999) (Jeffery, 2003) and a short protein form with trophic activity for cones respectively. Like RdCVF, RdCVF2 is transcribed into two spliced variant mRNAs, a long RdCVF2-L and a short RdCVF2-S forms (see Figure 87) translated into a long protein and a short protein shown to be involved in cone viability (Chalmel *et al., manuscript in preparation*).



	L	S
5'UTR	22 bp	44 bp
CDS	654 bp	330 bp
3'UTR	2076 bp	798 bp

	L	S
5'UTR	299 bp	264 bp
CDS	471 bp	306 bp
3'UTR	397 bp	2334 bp

Figure 87. RdCVF and RdCVF2 gene structure.
RdCVF and RdCVF2 gene structures are conserved. The gene structure and mRNA structures of RdCVF and RdCVF2 are presented in panels a and b respectively. Coding and non-coding regions are shown in dark grey and light grey respectively.

The RdCVF-L mRNA (NM_145598, mouse chromosome 8, minus strand, from 70,033,763 to 70,027,717) is composed of three exons (1-3) of 348, 687 and 1751 bp. The RdCVF-S mRNA (BC017153, from 70,033,785 to 70,032,615) is composed of a single exon (1,172 bp). The RdCVF2-L mRNA (AK015847, mouse chromosome 13, plus strand, from 50,202,630 to

50,206,797) is composed of two exons (1-2) of 603 and 564 bp. The RdCVF2-S mRNA (BC016199, from 50,202,667 to 50,205,571) is composed of a single exon (2,904 bp).

The mRNA structures reveal that the 5'UTR region of RdCVF2 is much larger than the 5'UTR of RdCVF. The CDS sizes of the short forms are comparable while the CDS of RdCVF-L is larger than the CDS of RdCVF2-L. The 3' UTR of RdCVF-L is significantly larger than that of RdCVF2-L one, due to a supplementary non-coding exon in RdCVF-L mRNA. Finally, the 3'UTR of RdCVF2-S is much longer than the 3'UTR of RdCVF-S.

The knowledge of these gene structures is essential to study the localization of the regulatory regions which are responsible for their expression regulation. The PromAn analyses for both of these two genes are presented in the subsequent sections.

11.3 Promoter analysis with PromAn

11.3.1 RdCVF

The genomic sequence from - 10 kb to + 10 kb relative to the whole RdCVF gene was studied with PromAn using the mouse sequence as reference (26,047 bp long) based on a TSS location at the position 10,001. Nine organisms were taken into account in this analysis: human, macaque, cow, dog, rat, opossum, chicken and frog (*Xenopus tropicalis*). This analysis was performed with the local version of PromAn and the TBA program was used to construct the multiple alignment.

The resulting profiles can be displayed with the PromAn GUI as illustrated in Figure 88.

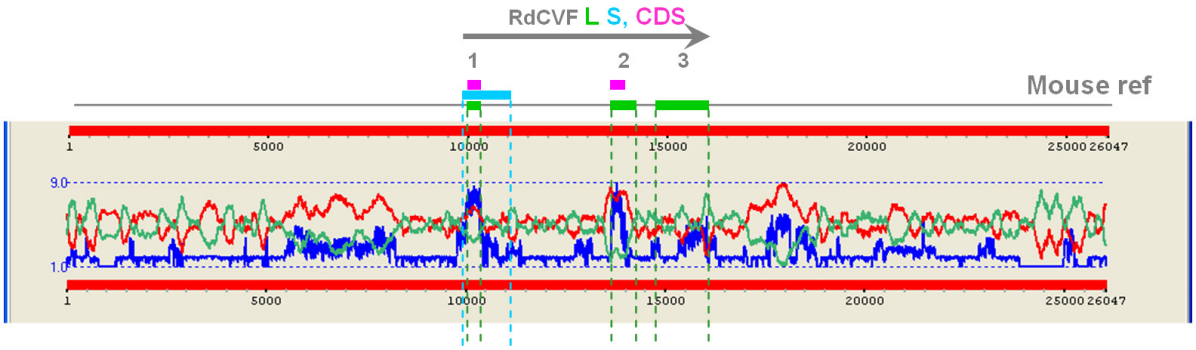


Figure 88. PromAn profiles of the RdCVF genomic sequence.
The mouse genomic sequence is used as reference (Mouse ref). The GC dinucleotide profile is shown in red and the AT dinucleotide profile in green. The conservation profile (*n*) (see section 6.1.2.3.2) is shown in blue. The RdCVF gene structure is shown above the PromAn profiles. The grey line is the mouse reference genomic sequence. The green boxes represent the location of the three exons of the long RdCVF mRNA (RdCVF-L), while the blue one corresponds to the short form. The coding region is represented with pink boxes.

The dinucleotide profiles of the RdCVF genomic sequence shows CpG islands as the brain tissue specific genes (see section 3.2.3). It is well known that coding regions are mostly well conserved during evolution. This is illustrated in PromAn by the fact that the highest peaks of conservation mainly correspond to coding exons. Thus, the two peaks highly conserved between the nine organisms, from human to chicken, correspond to the coding regions of the mouse RdCVF gene (displayed in pink). The non-coding part of the exons 1 and 2 and the non-coding exon 3 are less conserved. Note that the 3' part of the non-coding third exon shows a strong conservation in the 5 mammals studied. As frequently observed, the intronic regions are poorly conserved.

Interestingly, the conservation profile reveals other conserved regions upstream and downstream of the RdCVF exons which are conserved between mammalian organisms. The region immediately upstream of the RdCVF TSS shows a high conservation which quickly disappears and a second region of conservation (roughly from position 5,500 to 8,200). Similarly, several regions of high conservation are observed downstream of the third exon of the RdCVF gene, notably from position 17,500 to position 19,000.

The results validating the TSS position (see section 6.1.2.2.2) are shown in Figure 89.

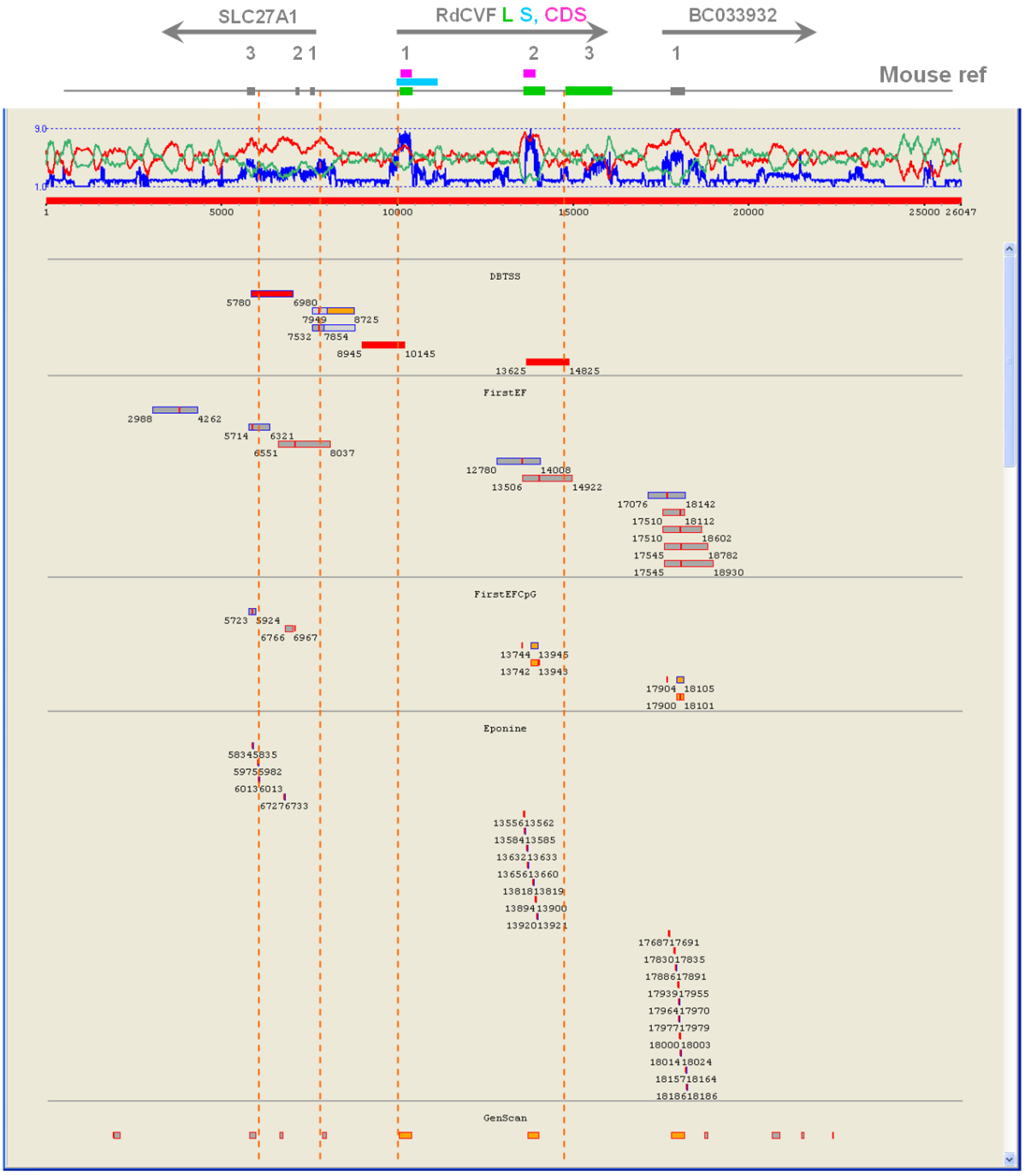


Figure 89. PromAn TSS location validation for the RdCVF mouse gene.

The results displayed in Figure 89 illustrate the heterogeneity of the predicted TSS. We observe that several experimental DBTSS promoters are located on the mouse reference sequence. PromAn reports that the DBTSS located between the positions 5,780 and 6,980 is described as the alternative promoter 1 of NM_011977 in DBTSS, and the two others shown in grey and orange are annotated as alternative promoter 2 of the same mRNA. The RefSeq mRNA NM_011977 corresponds to the SLC27A1 gene, annotated as a solute carrier family 27 (fatty acid transporter), member 1. It should be noted that the DBTSS alternative promoter 2 of NM_011977 has about 100 bp which differ from the mouse genomic sequence,

corresponding to a sequence error in this DBTSS promoter sequence. This difference of sequence results in two predictions (grey and orange) taking into account the size of the sequence which aligns with the mouse reference sequence. Nevertheless, these results strongly suggest that the conserved region from 5,500 to 8,200 corresponds to the SLC27A1 gene.

The DBTSS database provides the location of the experimental TSS as the position 9,945 (10,145-200), i.e. 56 bp away from the initial potential TSS at position 10,001. This promoter is annotated as the alternative promoter 1 of NM_145598. The RefSeq mRNA NM_145598 corresponds to the RdCVF-L mRNA. Interestingly, a second DBTSS promoter is highlighted with a TSS given at position 14,625 (14,825-200). This promoter is annotated as the alternative promoter 2 of NM_145598. It is located in the second intron at 165 bp from the non-coding exon (14,790-14,625). This observation has not been described before. Interestingly, the 3' end of the second intron (located immediately upstream of the non-coding exon) is conserved during evolution thus supporting the hypothesis of a functional TSS. This TSS could highlight the existence of a non-coding RNA which could for example be involved in the expression regulation of the RdCVF gene or most probably, of the downstream gene (BC033932) (see below) which is located close (1.8 kb) to the RdCVF gene, a situation known to be suitable for non-coding RNA regulation (see 2.2.2.2).

The promoter and TSS prediction programs FirstEF and Eponine predict wrong TSSs and promoters for the RdCVF gene. They approximately predict the TSS position for the SLC27A1 gene. They also predict the presence of a potential TSS in a third region located around position 18,000. Interestingly, this region (from position 17,500 to 19,000) is conserved during evolution. Taken together, these predictions strongly suggest the presence of a potential promoter and TSS in this region. Indeed, this hypothesis was verified through a GenBank BLAST search which revealed that the BC033932 mRNA has its first exon located in this region. This mRNA corresponds to an unannotated protein-coding gene. Finally, the GenScan program correctly predicts most of the known exons, thus allowing a correct indirect delineation of the potential TSS and promoter region.

These three genes are highlighted with arrows on the Figure 89 showing their respective location and sense. This example highlights the importance of considering a large genomic region for *in silico* promoter analysis, as here for example from -10 kb to + 10 kb related to the whole RdCVF gene. Thus, this large genomic region centered on the RdCVF gene takes into account the genomic context of three very close genes in the subsequent analysis and allows to hypothesise on the RdCVF gene expression regulation.

To study the potential regulatory regions of the RdCVF gene, we focus on the sequence ranging from the start codon of the SLC27A1 gene (position 5,880) to the start codon of BC033932 (position 17,793) (Figure 90).

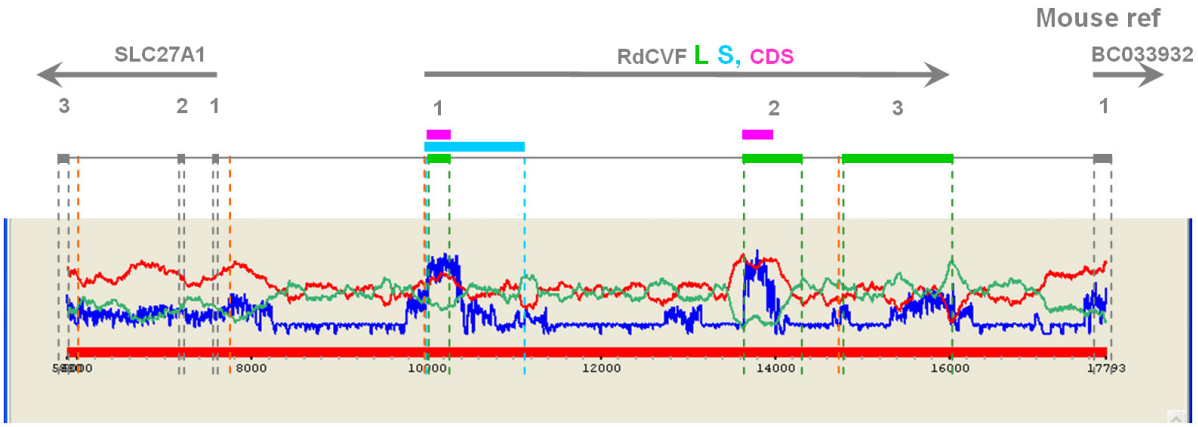


Figure 90. PromAn conservation profile of the region from the start codon of the SLC27A1 gene to the start codon of BC033932.

Preliminary analysis of this region clearly highlights that the RdCVF genomic region is particularly complex. Firstly, the 3 genes exhibit extremely short intergenic regions (from 1.7 to 2 kb) with respect to the average mouse intergenic regions. Secondly, with the notable exception of the 2 coding exons of the RdCVF gene, no strongly conserved region can be defined, either in intronic or intergenic regions or in the exons of the neighbouring genes. The only region presenting a slightly conserved segment is located in the 3' terminal part of the non coding exon of the long form of the RdCVF gene and might correspond to a non-coding RNA regulation previously discussed. Nevertheless, based on the conservation profile and on the genomic context we can suggest several hypotheses concerning to the regulation of the RdCVF gene expression. First, based on the small size of the intergenic sequence (2 kb) between the SLC27A1 and RdCVF genes and the fact that these two genes are present on opposite strands, we can suggest that this intergenic region may constitute a bidirectional promoter regulating the expression of both genes (see section 3.5). Second, if the promoter is not bidirectional, the conserved region of roughly 200 bp immediately upstream of the TSS of the RdCVF gene could be the proximal promoter responsible for the regulation of the RdCVF gene expression. Furthermore, regulatory elements could also be located in the conserved regions in the non-coding parts of the RdCVF exons or in the small slightly conserved regions present in the introns. Finally, the RdCVF gene expression could also be regulated by a potential non-coding mRNA which could have its TSS around the beginning of the third exon as presented above.

11.3.2 RdCVF2

The genomic sequence of the RdCVF2 gene was choosen from - 10 kb to + 10 kb relative to the start codon and analysed with PromAn using the mouse sequence as reference. Nine organisms were taken into account in this analysis: human, chimp, macaque, cow, rat, mouse, opossum, chicken and frog (*Xenopus tropicalis*). This analysis was performed with the local version of PromAn and the TBA program was used to construct the multiple alignment.

The resulting profiles can be displayed with the PromAn GUI as illustrated in Figure 91.

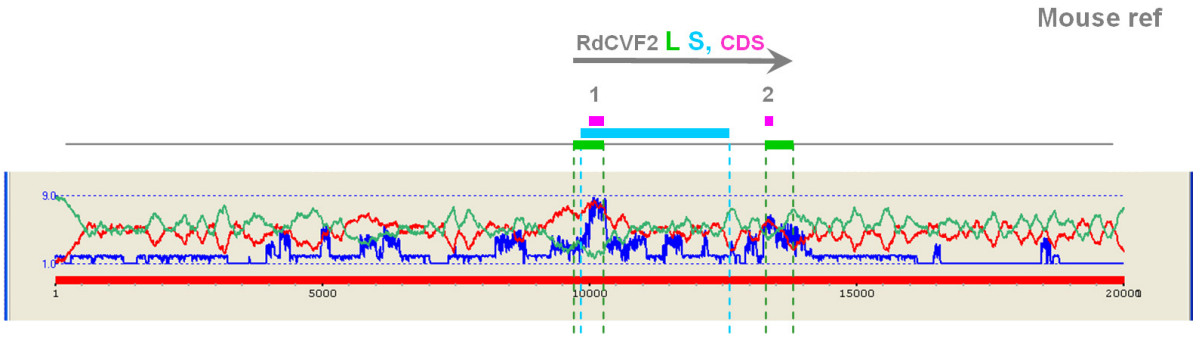


Figure 91. PromAn profiles from the analysis of the RdCVF2 genomic sequence.
The mouse genomic sequence is used as reference (Mouse ref). The GC dinucleotide profile is shown in red and the AT dinucleotide profile in green. The conservation profile (n) (see section 6.1.2.3.2) is shown in blue. The gene structure is presented above the PromAn profiles. The grey line is the mouse reference genomic sequence. The green boxes represent the location of the two exons of the long RdCVF2 mRNA (RdCVF2-L) while the blue one corresponds to the short form. The coding region is represented with pink boxes.

The conservation profile shows small peaks of conservation which are mostly conserved among mammals. The higher peaks of conservation strongly correlate with the first and second exons. In the 5'UTR regions, the conservation decreases with the distance from the start codon while the 3'UTR of the long mRNA form corresponds to a well conserved region. Finally, in the region located between the 2 exons and corresponding roughly to the 3'UTR region of the short mRNA form, some punctual conserved regions can be distinguished. Several peaks of conservation are also observed in the region upstream of the TSS.

The exact prediction of the TSS position (see section 6.1.2.2.2) is shown in the Figure 92.

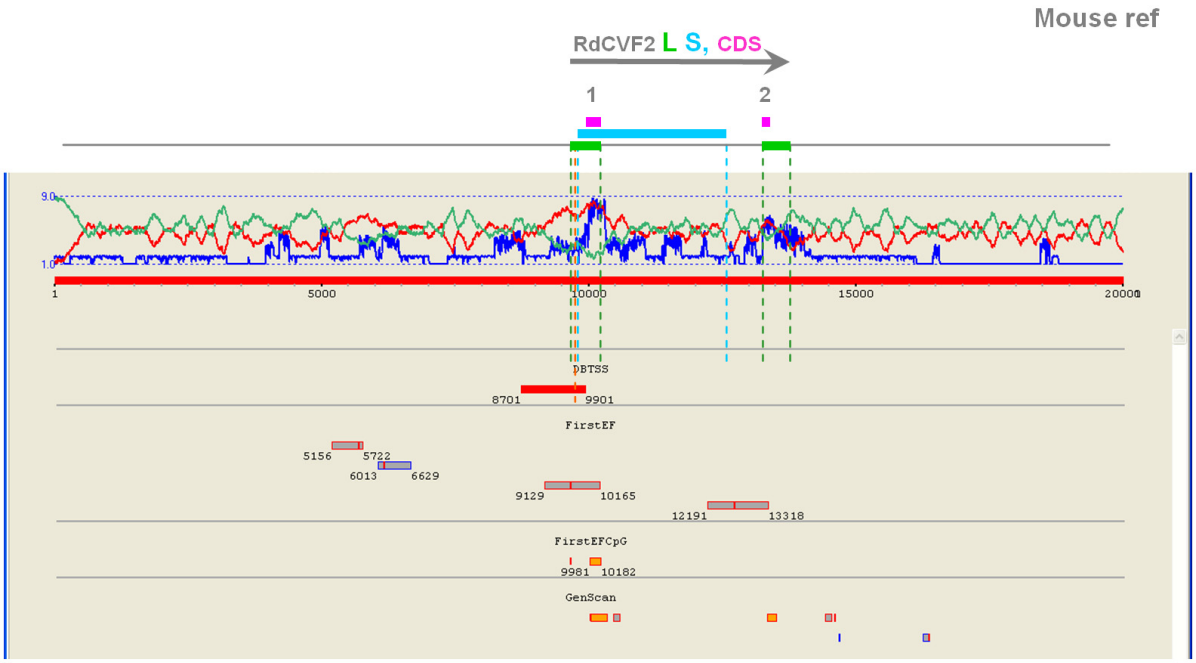


Figure 92. PromAn TSS location validation for the RdCVF2 mouse gene.

The DBTSS database locates the TSS of the RdCVF2 gene at position 9,701. The FirstEF program predicts the TSS region correctly, but also predicts other promoters downstream and upstream of the real promoter. Finally, the GenScan program predicts the two exons of RdCVF2 accurately, but also predicts supplementary exons, i.e. the real RdCVF2 TSS and exons are predicted but false positive exons are also predicted.

The regions of conservation observed from position 4,000 to the start codon constitute five potentially active promoter blocks. The first one corresponds to the proximal region up to -0.8 relative to the start codon. The other four regions of conservation are located approximately from -1.2 to -1.8, from -3.5 to -4.3 kb and from -4.8 to -5 and -5.6 to -6.1 kb relative to the start codon.

The experimental validations of the RdCVF and RdCVF2 genes are presented in the subsequent section.

11.4 Experimental validation

11.4.1 Method

Several DNA fragments corresponding to several putative promoters downstream of the start of translation were used as putative human and mouse promoters. The DNA fragments are ligated into the pGL4.17 reporter vector upstream of the easily assayed luciferase reporter gene. Each vector is then transfected into Y79 cells and expression of the reporter

gene is assayed. The same experiment was performed in parallel with the pGL4.17 vector alone.

The fold induction was calculated as the luminescent readings obtained with the pGL4.17 reporter plasmid containing a putative RdCVF(2) promoter compared with pGL4.17 reporter plasmid, given as:

$$Fold\ induction = \frac{\text{expression of pGL4.17 reporter plasmid with a putative RdCVF(2) promoter}}{\text{expression of pGL4.17 reporter plasmid}}$$

In the subsequent sections, we describe the putative promoter sequences tested and the results obtained. The putative promoter regions sometimes overlap with conserved regions highlighted with PromAn. This is due to the fact that these sequences were determined based on conserved regions calculated from a multiple genomic alignment performed with the multiz program. The conserved regions obtained with the TBA program are larger. Thus, the experimental validations will help to evaluate the quality of these alignment programs.

11.4.2 RdCVF

The putative mouse promoters tested are described in Table 23.

	Mouse potential RdCVF promoters				
Distance in kb related to the start codon	-4.2	-1.8	-0.9	-0.4	-0.13
Corresponding region on the PromAn mouse reference sequence: roughly from the following positions to the start codon	5,800	8,200	9,100	9,600	9,890

Table 23. RdCVF mouse genomic regions tested for their promoter activity.

The rationale for the DNA fragment design for the mouse promoter analysis (Figure 90) can be summarized as follows:

- The region of 4.2 kb corresponds to the genomic sequence from the start codon of the SLC27A1 gene to the start codon of the RdCVF gene. This construction was done under the assumption that a bidirectional promoter could regulate the expression of both SLC27A1 and RdCVF genes. This hypothesis is based on the small distance separating the TSS of these both genes (2 kb).
- The region of 1.8 kb only contains the region immediately upstream the RdCVF start codon as conserved. This construction should indicate if regulatory elements

present in the conserved region immediately upstream the start codon of SLC27A1 gene regulate the expression of the RdCVF genes.

- The region of 0.9 kb has a deletion of half of the non-conserved region of the construction of 1.8 kb.
- The 0.4 kb region has a deletion of three quarters of the non-conserved region of the construction of 1.8 kb and preserves the 0.28 kb conserved region located immediately upstream of the RdCVF gene.
- The 0.13 kb region corresponds approximately to half of the RdCVF proximal 0.28 kb conserved region.

Putative promoter regions of 2 kb, 1 kb and 0.55 kb relative to the human RdCVF start codon were also tested. As the human TSS is located at position 10,001 and the 5'UTR composed of 47 bp, the start codon is located at position 10,048 on the human genomic sequence used for PromAn analysis.

The putative human promoters tested are described in Table 24.

	Human potential RdCVF promoters		
Distance in kb related to the start codon	-2	-1	-0.55
Corresponding region on the PromAn human sequence: roughly from the following positions to the start codon	8,000	9,000	9,450

Table 24. RdCVF human genomic regions tested for their promoter activity

The rationale for human DNA fragment design has been defined based on the PromAn analysis of the RdCVF gene with the human genomic sequence as reference (*data not shown*). In the region of -2 kb relative to the TSS, only the region of -0.28 from the start codon presented a high conservation. Therefore, all the constructions tested contain this region of conservation that constitutes the putative proximal promoter.

The fold induction results are shown in Figure 93 and Figure 94 for the mouse and human RdCVF promoters respectively.

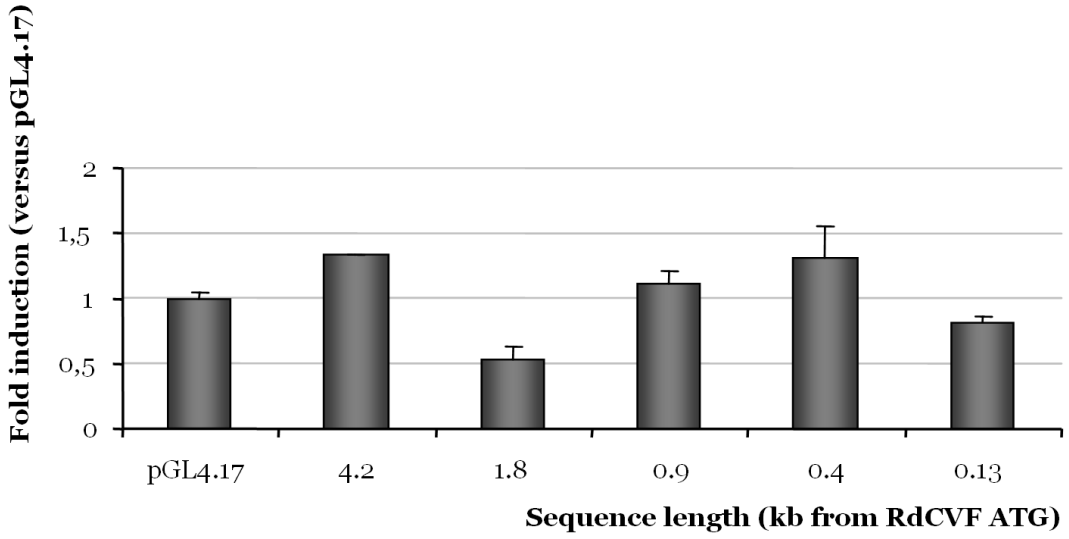


Figure 93. Mouse RdCVF

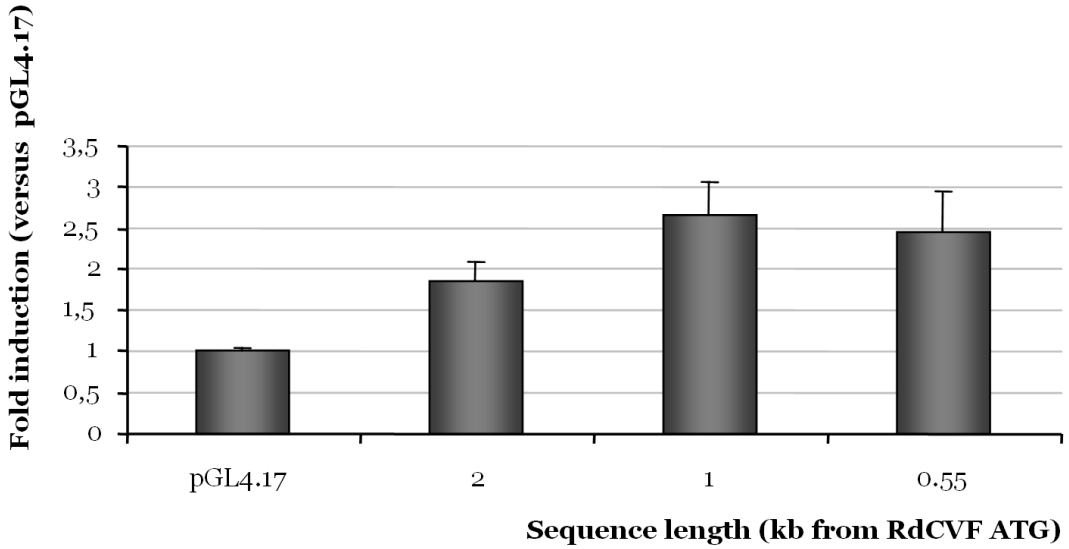


Figure 94. Human RdCVF

The mouse promoter sequences tested do not present a significant promoter activity in this current experiment. Indeed, the fold induction values obtained for the different RdCVF putative promoter regions tested do not significantly differ (highest luminescent reading of 1.4) from the pGL4.17 plasmid fold induction which is equal to 1. This experiment will be confirmed in the future.

The human promoters tested have an induction fold slightly higher than the pGL4.17 plasmid (from 1.8 to 2.7 compared to 1). These results suggest that regulatory elements are present in the human genomic sequences tested and that the human promoter is still present

in the region of 0.55 from the RdCVF start codon. Thus, this region would have to be cut more finely to refine these results and to localize more precisely the regulatory elements.

11.4.3 RdCVF2

The putative mouse promoters tested are described in Table 25.

	MousePotential RdCVF2 promoters			
Distance in kb related to the RdCVF2 gene start codon	-4.5	-2	-1.3	-0.9
Corresponding region on the PromAn human reference sequence: roughly from the following positions to the RdCVF2 start codon	5,800	8,300	9,000	9,400

Table 25. RdCVF2 mouse genomic regions tested for their promoter activity

The rationale for the DNA fragment design for the mouse promoter analysis (see Figure 91) is:

- The region of -4.5 kb relative to the start codon contains 3 regions of conservation previously described with the conservation profile.
- The region of -2 kb contains two regions of conservation.
- The region -1.3 kb contains the region of conservation immediately downstream of the start codon and the beginning of the second region of conservation.
- The region -0.9 only contains the proximal region of conservation.

The putative human promoters tested are described in Table 26.

	Human potential RdCVF2 promoters				
Distance in kb related to the RdCVF2 gene start codon	-4.5	-3.1	-2.1	-0.9	-0.26
Corresponding region on the PromAn human reference sequence: roughly from the following positions to the RdCVF2 start codon	5,800	7,200	8,200	9,400	10,071

Table 26. RdCVF2 human genomic regions tested for their promoter activity

The rationale for DNA fragment design has been defined according to a PromAn analysis performed on a human genomic region as reference (*data not shown*). The difference between the -4.5 and -3.1 genomic sequences corresponds to non-conserved regions. The construction of -2.1 overlaps half of the second region of conservation. The construction -0.9 only contains the proximal region of conservation, while the construction of -0.26 kb relative to the start codon only contains the 5'UTR region.

The fold induction results are shown in Figure 95 and Figure 96 for the mouse and human RdCVF promoters respectively.

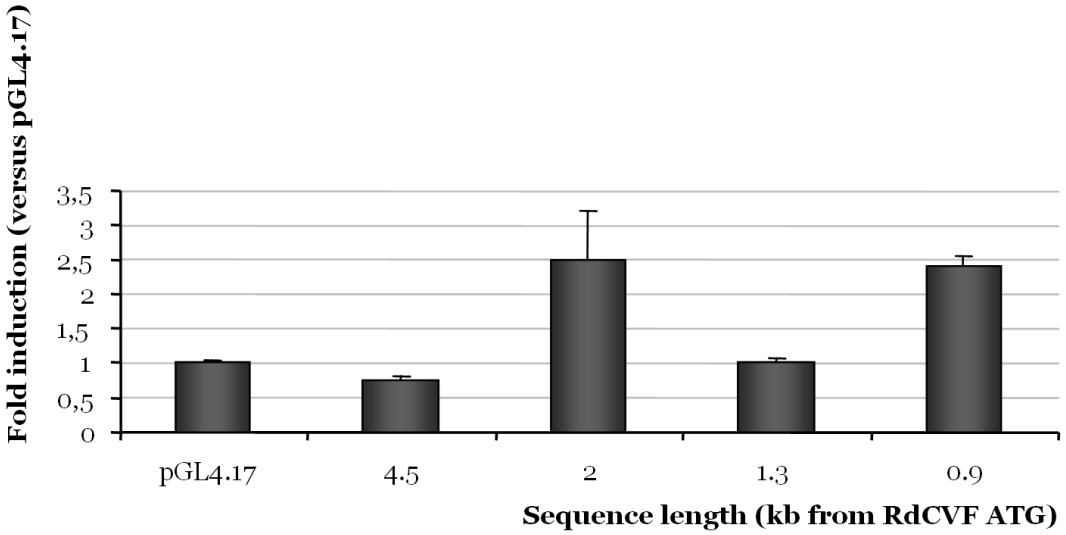


Figure 95. Mouse RdCVF2

The mouse promoter sequences tested do not present a promoter activity in the current experiment. Indeed, the fold induction values obtained for the different RdCVF2 putative promoter regions tested do not significantly differ from the pGL4.17 plasmid fold induction which is equal to 1 (comparison of the same luminescent readings). As this experiment was only performed once, it will be repeated before interpreting the results.

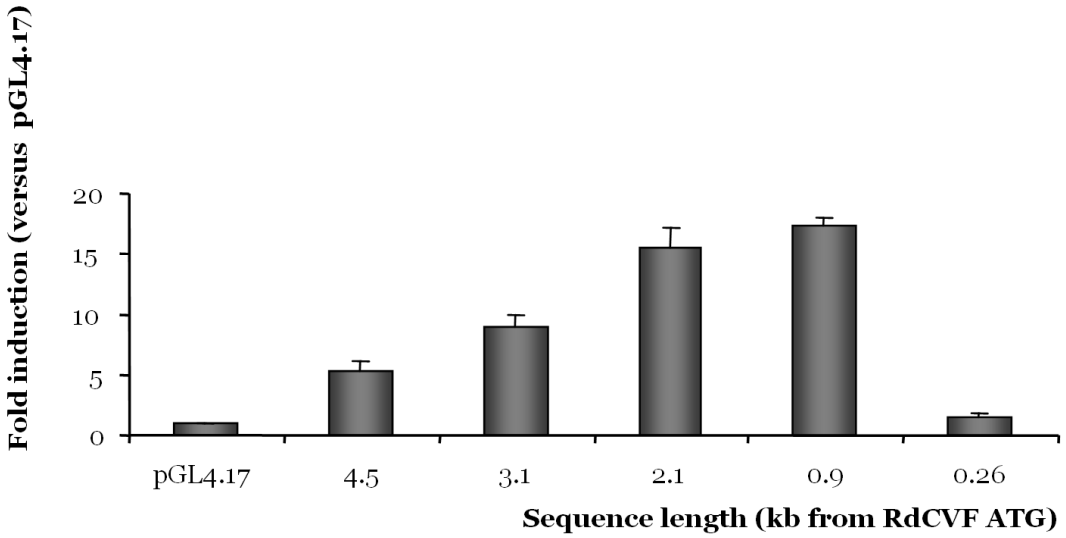


Figure 96. Human RdCVF2

The human RdCVF2 sequences tested present a promoter activity up to a fold induction of 18 compared with the pGL4.17 which has a value of 1. These results were confirmed with

several experiments in different cell types. Therefore we conclude that in the RdCVF2 human gene, the region between -0.26 and -0.9 kb presents a promoter activity. The proximal conserved region downstream of the TSS plays a role in the activation of the RdCVF2 gene expression. Furthermore, for the constructions from -4.5 kb to -0.9 kb, we observe an increased fold induction. This may highlight either the presence of silencing elements in this distal region which would be mostly unconserved during evolution, or also it may be explained by the transfection efficiency which could decrease with the increased plasmid size.

In conclusion, further experiments are now in progress to confirm the results obtained for the human and mouse RdCVF gene and for the mouse RdCVF2 gene. The analysis of the RdCVF2 human promoter is also in progress. Precise analysis of the multiple alignment in the region from -0.26 to -0.9 is in progress to further orientate experiments aimed at refining the regulatory regions for the identification of the regulatory CRMs (*cis* regulatory modules). Identification of the potential TFBSs present in the region from -0.26 to -0.9 kb relative to the human RdCVF2 start codon are also in progress and will be experimentally tested to identify the TFs which are responsible for the regulatory activity of this region. Classical methods such as EMSA, footprinting and siRNA to block the translation of a dedicated TF will be used.

Chapter 12 - Analysis of neurodegenerative diseases caused by polyglutamine disorders.

12.1 Scientific context

Huntington's disease (HD) and spinocerebellar ataxia type 7 (SCA7) belong to a group of inherited neurodegenerative diseases classified as polyglutamine (polyQ) disorders. The causal mutation is a CAG repeat expansion in the corresponding genes encoding an expanded polyQ tract in the proteins. PolyQ expansions in mutant proteins confer toxic properties such as aberrant interactions with normal protein partners and accumulation in neurons to form intranuclear inclusions (NIs), a microscopic hallmark of these diseases (Ross, 2002). Although polyQ disorders share common genetic features, they generate distinct patterns of neuronal degeneration despite overlapping protein expression areas (Zoghbi and Orr, 2000). For example, the primary target in HD is the striatum, whereas SCA7 causes degeneration of the cerebellum and brain stem and is the only polyQ disorder affecting the retina. Transcriptional alteration is a unifying feature of polyQ disorders; however, the relationship between polyQ-induced gene expression deregulation and degenerative processes remains unclear. R6/2 and R7E mouse models of HD and SCA7, respectively, present a comparable retinal degeneration characterized by progressive reduction of electroretinograph activity and important morphological changes of rod photoreceptors.

At the physiological level: until 3 weeks of age, R7E retina develops normally and displays no obvious phenotype. At 3 weeks, NIs are detected in some rod photoreceptors and the first functional abnormalities detected by scotopic electroretinogram (ERG) occur between 3 and 4 weeks of age. At 9 weeks, the moderate stage is characterized by the presence of NIs in most rod cells and a marked reduction of ERG response. The nine-week-old retina also displays morphological alterations of rod photoreceptors characterized by loss of segments and enlarged nuclei with atypical decondensed chromatin (Helmlinger et al., 2004) (Helmlinger et al., 2006). However, there is no significant neuronal loss at this stage. Later on, R7E retinopathy worsens towards flattening of ERG recording, complete loss of segment layers and thinning of the outer nuclear layer.

The analysis performed in collaboration with Gretta Abou-Sleymane in the laboratory of Yvon Trottier at the IGBMC was aimed at elucidating the molecular and cellular events underlying retinal degeneration in R7E and R6/2 mice (Abou-Sleymane et al., 2006) (see Annexe 1 -).

12.2 Transcriptomic analysis

To correlate gene deregulation during the progression of R7E retinal degeneration, expression profiling analysis of R7E and R6/2 retina was performed. R7E and the control R7N mice gene expression profiles were examined at onset (3 weeks of age) and moderate (9 week of age) stage of pathology with Affymetrix microarrays (MOE430A). The same analysis was performed with the R6/2 versus the wild type.

The initial data preparation was pre-processed and normalized by MAS5 (MicroArray Suite Version 5.0) using Affymetrix default analysis settings and global scaling as the normalization method. The mean intensity of each chip was arbitrarily set to 100. Absolute analyses generate a signal value for each probe set and a detection call of "absent", "present" or "marginal". To select differentially expressed genes, three consecutive filters were applied for each of the three comparison groups (3-week-old R7E versus R7N, 9-week-old R7E versus R7N and 9-week-old R6/2 versus wt). First, a Mann-Whitney statistical test was performed using a *p-value* threshold < 0.015. Then, genes presenting an Affymetrix fold change (AFC) >1.5 were selected. Finally, only genes called "present" in at least (n - 1) mice in a group of n mutant mice for the up-regulated genes, or n control mice for the down-regulated genes were selected.

The MOE430A Affymetrix GeneChip contains 22,600 probe sets. About 15,000 of them correspond to known genes, whereas the remaining probe sets correspond to EST sequences. Of the 22,600 probe sets, the comparison of gene expression of R7E versus R7N led to 45 and 61 transcripts over- and under-expressed respectively at the onset stage, while 185 and 301 transcripts were over- and under-expressed at the moderate stage. Retinas of moderate R6/2 mice were compared to wt littermate mice. Of the 311 differentially expressed transcripts, 81 were identified as over-expressed and 230 as under-expressed.

12.3 Automatic transcript annotation using the RetScope platform and GOAnno.

Because of the large number of unknown transcripts (33.6%) or poorly characterized genes, we used a systematic approach to annotate differentially expressed transcripts using RetScope, an automatic sequence analysis platform developed in our laboratory. Briefly, RetScope, associates with high confidence each probe set sequence with existing full-length transcript and protein accession numbers. This first annotation step is based on the analysis of BLAST (Altschul et al., 1990) homology searches in the GenBank (Benson et al., 2005), UniProt (Bairoch et al., 2005) and Human genome (Lander et al., 2001) public databases.

Each protein sequence related to the differentially expressed transcript was then automatically annotated with GO terms using the GOAnno program (Chalmel et al., 2005) (see Chapter 7 -). Of the 727 non-redundant transcripts in the 6 clusters, 541 could be assigned to "biological process" terms.

12.4 GO term enrichment calculation for co-expressed genes

We identified the "biological process" terms enriched in the six described groups of co-regulated genes (up- and down-regulated in R7E onset, R7E moderate and R6/2 moderate).

GO terms were assigned as "significantly enriched" using the binomial statistical model with a *p-value* threshold of 0.00135 corresponding to a *z score* cutoff equal to 3. These enriched GO terms are shown in the Figure 97 for each group of co-regulated genes.

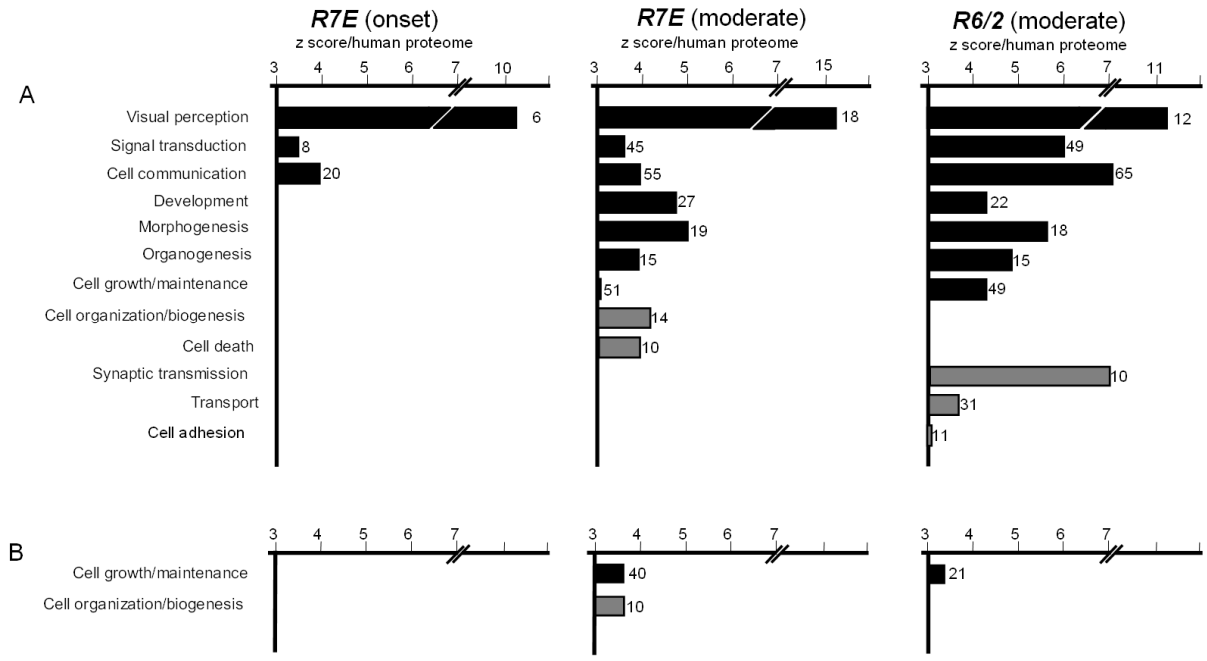


Figure 97. Enriched "biological process" GO terms corresponding to R7E and R6/2 de-regulated genes.
Panel A and B display the under- and over-expressed genes respectively in retina of R7E onset, R7E moderate and in R6/2 moderate. Only "biological process" terms presenting a *z score* value >3 and including at least six genes for R7E onset and 10 genes for R7E and R6/2 moderate are shown. Furthermore, the number of transcripts annotated with the corresponding GO term is indicated at the right side of each histogram bar. Black bars represent "biological process" GO terms common to R7E and R6/2, whereas gray bars are specific to a single mutant mouse.

Disease onset in R7E was associated with under-expressed genes significantly enriched in *signal transduction*, *cell communication* and most importantly *visual perception* processes (Figure 97). The number and enrichment of under-expressed genes of these three functional categories increased at the moderate stage of the R7E phenotype. In contrast to signal transduction and cell communication, visual perception represents a level of higher specificity according to the GO annotation system. Enrichment of down-regulated genes involved in visual perception makes perfect sense with the early and progressive ERG defect in R7E retinopathy and in itself validates the method that we used to identify deregulated pathways.

The moderate stage of the R7E phenotype was also characterized by enrichment of down-regulated genes involved in development, morphogenesis and organogenesis as well as up- or down-regulated genes associated with cell growth, maintenance and organization. Expression alteration of genes belonging to these categories might account for the important morphological changes (e.g. loss of segments and enlarged nuclei) of rod photoreceptors

seen at 9 weeks. Down-regulation of genes involved in regulation of cell death was also over-represented, although rod photoreceptor death is observed at later stages only.

Strikingly, altered genes in R6/2 were significantly enriched in the same functional categories as in R7E at 9 weeks. This concerned genes involved in visual perception, signal transduction, cell communication, as well as in *development*, *morphogenesis*, *organogenesis* and *cell growth/maintenance* (Figure 97). These results highlight at the molecular level the similitude of retinal degeneration and dysfunction in the two models. As might be expected, some genes were enriched in categories unique to R6/2. Indeed, under-expressed genes involved in synaptic transmission, transport and cell adhesion are likely to underlie retinal dysfunctions in R6/2 mice.

12.5 TFBS enrichment calculation for co-expressed genes

In this analysis, three key TFs (*Nrl, Crx* and *Nr2e3*) controlling rod differentiation genes, hence expression of photoreceptor specific traits, are down-regulated. Furthermore, other TFs known to cause inhibitory effects on photoreceptor differentiation when mis-expressed, such as *Stat3* (signal transducer and activator of transcription 3), are aberrantly re-activated. Thus, these results suggest that independently from the protein context, polyQ expansion may override the control of neuronal differentiation and maintenance, thereby causing dysfunction and degeneration.

As *stat3* is re-activated and has a strong transcriptional activation activity, we hypothesised that genes regulated by this TF should also be present in clusters of genes which are over-expressed (see Figure 97, panel B). These three clusters are non-photoreceptor specific. They contain genes which are not implicated in the vision biological process and are less well characterized in GO annotation. In order to test the hypothesis that these three clusters containing over-expressed genes would present enrichment in genes regulated by the *stat3* TF, we considered their enrichment in predicted *stat3* TFBSs. The *stat3* TFBS was search with the TRANSFAC matrix M00497. The high-throughput version of PromAn was used for this analysis (see section 6.3). A search was performed for the *stat3* motif in the genomic regions of the genes present in three selected clusters and in two random datasets (see section 8.3.2). Various parameters were used as the location relative to the TSS position (0/+200, -200/+200, -500/+200, -1,000/+200 and -2,000/+200) and ranges of matrix and conservation scores (from 0 to 1, from 0.5 to 1, from 0.8 to 1) were also tested. All the data obtained were used to calculate the *stat3* TFBS enrichment in each cluster and for the different conditions of location relative to the TSS and of matrix and conservation scores.

No significant enrichment (matrix and conservation scores >0.8 and *p-value* <0.05) in *stat3* TFBSs was observed in over-expressed genes. This analysis was performed under the assumption that the selected genes are likely to be implicated in common biological processes and thus could be regulated by common TFs. The three studied clusters contain co-expressed genes, nevertheless according to the GO study they are not enriched in common biological processes (see section 12.4) with the exception of the cluster R7E which presents enrichment in very global mechanisms (cell growth/maintenance, cell organization and biogenesis), that are less specialized than the visual perception processes found to be enriched in the three other clusters containing under-expressed genes.

To improve the enrichment estimation, a common approach is to refine the clustering for the selection of sub-groups of co-expressed genes among which an over-representation of *stat3* TFBSs may be observed. This could be done by adding other experimental conditions to improve the potential correlation between co-expression and co-regulation. In this case, we would need groups containing more genes in order to be able to calculate significant statistical enrichments.

Finally, it should be stressed that *stat3* may activate genes which correspond to other TFs and consequently, the clusters could present an enrichment in TFBSs of these TFs. This could be verified through additional analysis aiming at the identification of all TFs containing at least one *stat3* TFBS prediction in their promoter and present in the clusters of over-expressed genes. This would involve the calculation of a potential enrichment of their corresponding TFBS in the over-expressed set of genes.

This analysis characterizes co-expressed genes based on limited conditions (two temporal points) and using a single tissue as model (retina). Clearly, in view of the complex regulatory networks at work in cellular and development processes, these experimental conditions may not be sufficient to uncover the many distinct TFs and TFBSs that might be responsible for the observed correlated co-expression and co-regulation. We thus assume that co-expressed genes characterized with more experimental conditions (expression data in several tissues, or more temporal points for example) may improve the probability of co-regulation, allowing the reliable identification of TFBS enrichments. To verify this hypothesis, we realized further analyses of TFBS enrichment on distinct microarray analyses performed with more experimental conditions.

Chapter 13 - Analysis of transcriptomics data with HT PromAn

13.1 Scientific context

As illustrated previously, it is clear that in order to highlight a potential co-regulation by specific TFs in a group of genes, a stringent selection of co-expressed genes is necessary. As previously illustrated in the section 12.5 , the characterization of co-expressed genes based on limited temporal conditions does not appear to be sufficient for a statistically significant selection of genes potentially co-regulated by the same TF.

In the context of specialized biological processes in specialized tissues (hereafter, we will use as an example system, the establishment or maintenance of the layer of the photoreceptors in the retina or the differentiation of the male germ cells in the testis), we are interested in selecting genes preferentially expressed in a given cell type (rod photoreceptors or spermatids). As these selected genes can also be expressed in other organs (or tissues), we are interested in selecting those genes which are specifically expressed in one tissue (retina or testis). These steps of selection aim at filtering genes preferentially expressed in a given tissue or cell type only and thus at improving the signal to noise ratio. This approach is based on the assumption that those genes should be more rigorously co-regulated and thus should have very strong common signals in their promoter (see section 3.1). We hypothesise that these genes might be regulated by the same TFs which would be responsible for their specific expression profiles. It allows us to test our TFBS enrichment method (see section 8.3.2) on various groups of genes involved in very specific processes.

To this end, pre-processed and pre-analysed data from two microarray experiments were provided by the laboratory of Dr Thierry Léveillard (Laboratoire de Physiopathologie Cellulaire et Moleculaire de la Retine, Faculté de Médecine Saint-Antoine, Paris) and the laboratory of Dr Michael Primig (Biozentrum, University of Basel, Switzerland). The first project involves the kinetic study of retinal degeneration in the *rd1* mouse compared to the wild-type (wt) mouse and during the first post-natal (PN) month. The second project involves the expression program analysis of germ cells during the spermatogenesis process in mouse. Both of these studies will be described in detail in the following sections. Furthermore, the laboratory of Dr Michael Primig also provided transcriptomic data from 25

normal mouse tissues which were extracted from the NCBI GEO public repository (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) and further pre-processed and normalized. These expression data allowed us to perform a tissue profiling filtration to select retina or testis-specific genes.

13.1.1 *rd1* mouse transcriptome analysis

As described in Chapter 11 -, retinitis pigmentosa is a two-stage inherited retinal disease. Rod photoreceptors degenerate first, resulting in the loss of night vision, followed by the progressive loss of light-adapted vision, due to secondary cone photoreceptor death. The secondary cone death leads to blindness in middle age, as cones are essential for diurnal, color and high acuity vision. A large proportion of the retinitis pigmentosa mutations affect genes expressed primarily in rod photoreceptor cells (Phelan and Bok, 2000).

The retinal degeneration 1 (*rd1*) mouse, the most commonly studied model of retinitis pigmentosa carries a recessive mutation in the rod-specific cGMP phosphodiesterase beta subunit gene (Pde6b) leading to rod photoreceptor death through apoptosis (Portera-Cailliau et al., 1994), followed by cone death presumably through lack of trophic support from rod-derived cone viability factors (Mohand-Said et al., 1998). The precise molecular mechanisms leading from the increased cGMP concentration induced by the enzymatic defect to rod apoptosis remain unknown. Therefore, the *rd1* mouse is an appropriate model for the investigation of the transcriptional events in rod death and provides the opportunity to identify clusters of co-expressed genes that are related to different steps in the process of rod degeneration, including genes that might control apoptosis.

The laboratory of Dr Thierry Léveillard used a transcriptomic approach with the Affymetrix GeneChip mouse430_2 to identify genes and pathways that could be targeted for the treatment of retinitis pigmentosa (Chalmel *et al.*, *manuscript in preparation*). An expression profile comparison between wt and *rd1* mice was performed for 12 time points, from post-natal day 5 to 28, the period in which rod photoreceptors degenerate.

13.1.2 Transcriptome analysis of the mouse spermatogenesis

Meiosis and gametogenesis are critical processes in the transmission of genetic material to subsequent generations during sexual reproduction. In males, mitotically growing spermatogonia develop into meiotic spermatocytes that give rise to haploid spermatids which differentiate into mature sperm. This pathway is controlled in part by somatic Sertoli cells that physically interact with germ cells and communicate with them via hormonal cues. Many of the loci required for meiotic landmark events such as recombination and gamete

formation are expressed only in cells capable of undergoing the process but the regulatory network that confers germline-specific transcription in higher eukaryotes is only poorly understood, especially at the mitotic and meiotic stages (Maclean and Wilkinson, 2005). In this study, purified spermatogonia (mitotic germ cells), spermatocytes (meiotic) and spermatids (post-meiotic) were compared to somatic Sertoli cells to select differentially expressed loci. This study is aimed at providing a framework for the functional characterization and clinical application of numerous known and novel genes predicted to be involved in meiosis and gametogenesis in mammals (Chalmel *et al.*, *submitted*).

13.1.3 Tissue profiling transcriptomics data

To refine putative retina- or testis-specific gene lists, a representative set of data from 25 normal mouse tissues available via NCBI GeneOmnibus was assembled. This set includes one retina, one eye and one brain sample, six mouse control testicular samples (A- and B-type spermatogonia, pachytene spermatocytes, post-meiotic round spermatids and total testis), two female reproductive cell types (cumulus-oocyte complex and ovary) and 14 other tissue samples (placenta, aorta, heart, lung, kidney, liver, spleen, submaxilar gland, thymus, stomach, lymphoid node, skeletal muscle, CD4 naïve and embryo D7).

13.1.4 From raw data to clusters of potentially co-regulated genes

The microarray experiments, as well as the statistical filtering and the clustering of differentially co-expressed genes of both *rd1* and spermatogenesis transcriptomic projects are briefly described in Table 27.

	rd1 transcriptome	Spermatogenesis transcriptome
Organism	Mouse	Mouse
Strain	wt versus *rd1*	wt
Cell types or tissues	Total retina	Sertoli cells (SE), spermatogonia (SG), spermatocytes (SC), spermatids (ST), tubules (TU) and total testis (TT)
Stage or temporal points	12 post-natal days: 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 21 and 28	Adult
Replicates	Duplicates or triplicates per day and per strain	Duplicates per cell type
Affymetrix microarray	mouse430_2	mouse430_2
Number of probe sets	45101	45101
Method of statistical filtration of the probe sets	Fold-change > 2 between wt and *rd1* in one of the 12 days. F-value adjusted with FDR (False Discovery Rate) < 0.05	Standard deviation across samples > 1. Permutation Tests with *p-value* < 0.001
Number of differentially expressed probe sets	1834	9283
Clustering method	PAM (Partition Around Medoids)	PAM
Number of clusters	6: G1 to G6	4: Somatic (SO) ; Mitotic (MI) ; Meiotic (ME) ; Post-Meiotic (PM)

Table 27. *rd1* and spermatogenesis transcriptome analysis.

In the *rd1* transcriptome analysis, genes whose expression signals varied significantly during the first post-natal month between the *rd1* and wt strains were selected. This approach identified 1,834 differentially expressed probe sets representing approximately 4% of the transcripts present on the GeneChips. The supervised PAM (*Partition Around Medoids*) clustering algorithm was used to identify 6 groups of co-expressed genes (G1 to G6) corresponding to 6 distinct expression profiles. 1,834 probe sets were thus classified into (Figure 98):

- G1 (377 probe sets) shows an increasing expression profile in both wt and *rd1* strains but with a higher expression in *rd1* at PN day 28.
- G2 (344) shows a profile of increasing expression in wt during all the PN days. The *rd1* strain has the same expression profile until PN day 9 followed by a huge decrease of expression.
- G3 (189) displays a flat profile in the wt strain and an increasing profile in *rd1* strain.
- G4 (384) shows a flat profile in the wt while the *rd1* profile decreases from PN day 11.

- G5 (204) displays a decreasing profile in both wt and *rd1* strains but with a high expression in *rd1* at PN day 28.
- G6 (336) shows an increasing profile in wt until PN day 11 followed by a flat profile. In the *rd1* strain has the same expression profile unti PN day 13 followed by a huge decrease of expression.

In the wt strain, the G2 and G6 groups of co-expressed transcripts follow the expression profile of known rod photoreceptor marker genes such as the Gnat1 (Guanine nucleotide binding protein, alpha transducing 1) or Guca1a (Guanylate cyclase activator 1a) in G2 and Pde6b (rod phosphodiesterase 6b) or Rho (rhodopsin) in G6. These groups are representative of the rod photoreceptor genes that dramatically decrease with the rod degeneration at work in *rd1*. Thus, we will further focus our analysis on these two G2 and G6 clusters. These genes appear to be expressed in the photoreceptors (GO enrichment calculation will confirm this hypothesis, see below), nevertheless as observed in Figure 98 most of them are also expressed in other tissues.

In the spermatogenesis transcriptome analysis, genes whose expression signals varied significantly between Sertoli cells, germ cells, tubules and whole-organ controls were selected. This approach identified 9,283 probe sets corresponding to approximately 21% of the transcripts present on the GeneChips. The PAM clustering algorithm was used to separate mouse genes into four determined groups that best fit the cell populations analysed. 9,283 probe sets were thus classified into somatic (2,134 probe sets, SO), mitotic (3,243, MI), meiotic (2,317, ME) and post-meiotic (1,589, PM) expression clusters showing transcriptional peaks in somatic Sertoli cells, mitotic spermatogonia, meiotic spermatocytes and post-meiotic round spermatids, respectively (Figure 98). Note that this classification is based on the strongest level of induction that does not necessarily reflect cell-type specific expression.

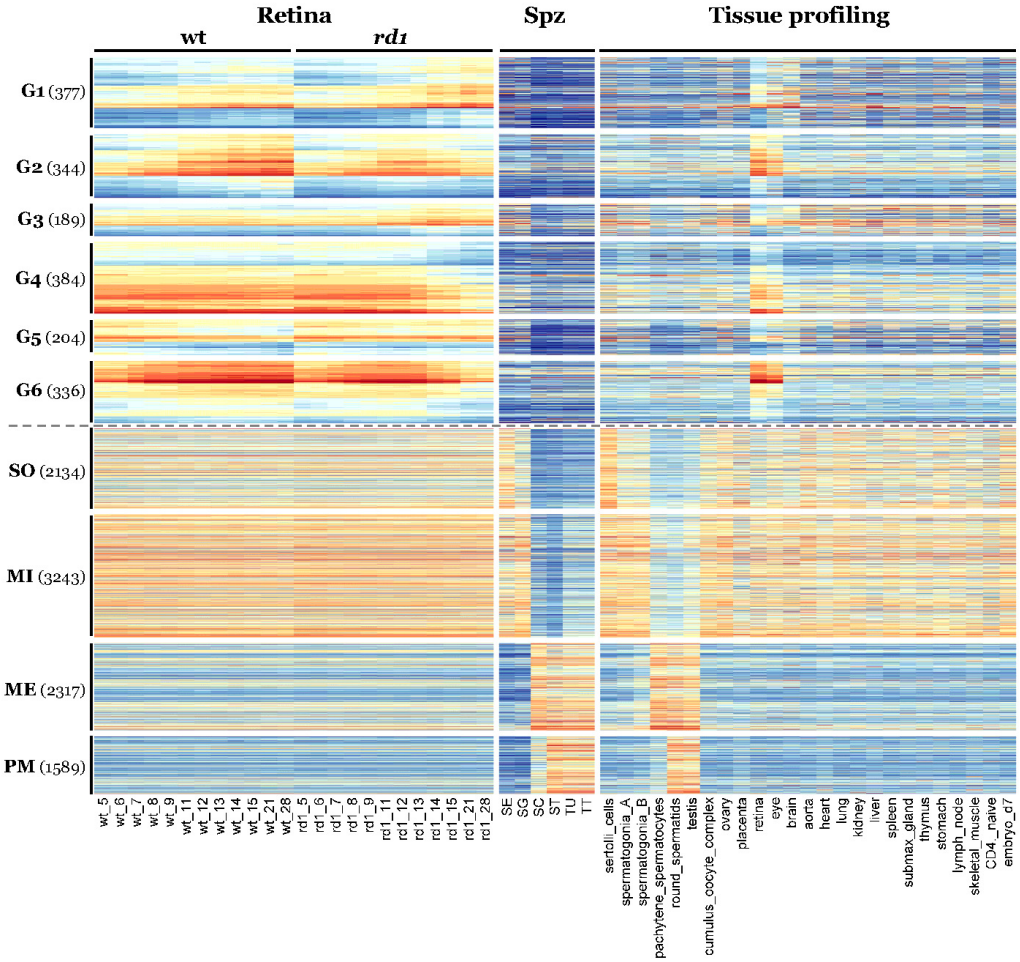
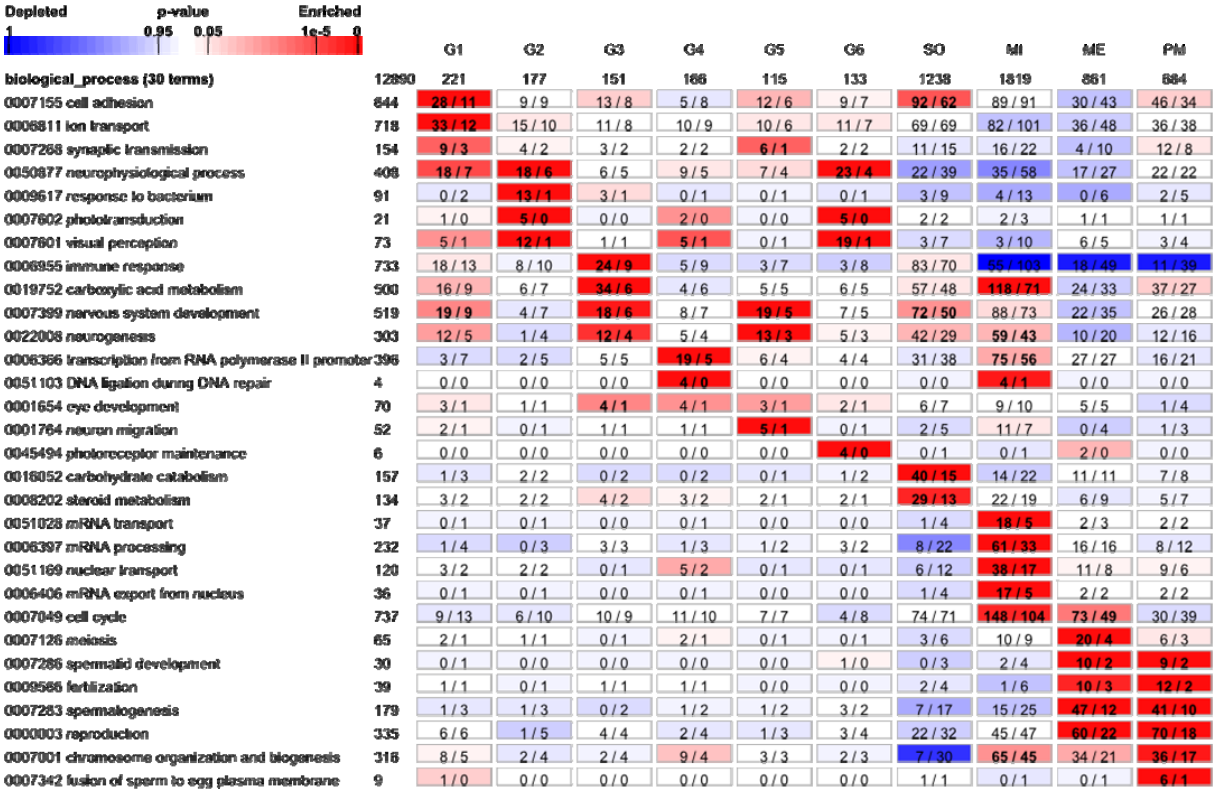


Figure 98. Heatmap of the selected groups of co-expressed genes in the *rd1* and wt retina, germ cells and tissue profiling transcriptomics data.
False color-coded heatmaps of the retina, spermatogenesis and of the tissue profiling studies are shown in the same figure. The six expression clusters (G1 to G6) of genes differentially expressed between the *rd1* and the wt mouse during the first PN month are shown above the grey dashed line while the four expression clusters (SO, MI, ME and PM) of genes differentially expressed between germ cells are shown below this dashed line. Each line represents a probe set corresponding to one gene and each column corresponds to the different experimental conditions previously described. SE, SG, SC, ST, TU and TT experimental conditions correspond to Sertoli cells, spermatogonia, spermatocytes, spermatids, tubules and total testis respectively. The names of the clusters and the numbers of probe sets in each expression cluster are shown on the left. Log2-transformed expression signals are color-coded from blue (low expression) to red (high expression).

The clusters of both projects were characterized with GO terms in order to highlight putative significantly enriched biological processes (Figure 99).



		G1	G2	G3	G4	G5	G6	SO	MI	ME	PM
biological_process (30 terms)	12890	221	177	151	166	115	133	1238	1819	861	684
0007155 cell adhesion	644	**28 / 11**	9 / 9	13 / 8	5 / 8	12 / 6	9 / 7	**92 / 62**	89 / 91	30 / 43	46 / 34
0006811 ion transport	718	**33 / 12**	15 / 10	11 / 8	10 / 9	10 / 6	11 / 7	69 / 69	82 / 101	36 / 48	36 / 38
0007268 synaptic transmission	154	**9 / 3**	4 / 2	3 / 2	2 / 2	**6 / 1**	2 / 2	11 / 15	16 / 22	4 / 10	12 / 8
0050877 neurophysiological process	408	**18 / 7**	**18 / 6**	6 / 5	9 / 5	7 / 4	**23 / 4**	22 / 39	**35 / 58**	17 / 27	22 / 22
0009617 response to bacterium	91	0 / 2	**13 / 1**	3 / 1	0 / 1	0 / 1	0 / 1	3 / 9	4 / 13	0 / 6	2 / 5
0007602 phototransduction	21	1 / 0	**5 / 0**	0 / 0	2 / 0	0 / 0	**5 / 0**	2 / 2	2 / 3	1 / 1	1 / 1
0007601 visual perception	73	5 / 1	**12 / 1**	1 / 1	**5 / 1**	0 / 1	**19 / 1**	3 / 7	3 / 10	6 / 5	3 / 4
0006955 immune response	733	18 / 13	8 / 10	**24 / 9**	5 / 9	3 / 7	3 / 8	83 / 70	**55 / 103**	**18 / 49**	**11 / 39**
0019752 carboxylic acid metabolism	500	16 / 9	6 / 7	**34 / 6**	4 / 6	5 / 5	6 / 5	57 / 48	**118 / 71**	24 / 33	37 / 27
0007399 nervous system development	519	**19 / 9**	4 / 7	**18 / 6**	8 / 7	**19 / 5**	7 / 5	**72 / 50**	88 / 73	22 / 35	26 / 28
0022008 neurogenesis	303	12 / 5	1 / 4	**12 / 4**	5 / 4	**13 / 3**	5 / 3	42 / 29	**59 / 43**	10 / 20	12 / 16
0006366 transcription from RNA polymerase II promoter	396	3 / 7	2 / 5	5 / 5	**19 / 5**	6 / 4	4 / 4	31 / 38	**75 / 56**	27 / 27	16 / 21
0051103 DNA ligation during DNA repair	4	0 / 0	0 / 0	0 / 0	**4 / 0**	0 / 0	0 / 0	0 / 0	**4 / 1**	0 / 0	0 / 0
0001654 eye development	70	3 / 1	1 / 1	**4 / 1**	**4 / 1**	3 / 1	2 / 1	6 / 7	9 / 10	5 / 5	1 / 4
0001764 neuron migration	52	2 / 1	0 / 1	1 / 1	1 / 1	**5 / 1**	0 / 1	2 / 5	11 / 7	0 / 4	1 / 3
0045494 photoreceptor maintenance	6	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	**4 / 0**	0 / 1	0 / 1	2 / 0	0 / 0
0016052 carbohydrate catabolism	157	1 / 3	2 / 2	0 / 2	0 / 2	0 / 1	1 / 2	**40 / 15**	14 / 22	11 / 11	7 / 8
0008202 steroid metabolism	134	3 / 2	2 / 2	4 / 2	3 / 2	2 / 1	2 / 1	**29 / 13**	22 / 19	6 / 9	5 / 7
0051028 mRNA transport	37	0 / 1	0 / 1	0 / 0	0 / 1	0 / 0	0 / 0	1 / 4	**18 / 5**	2 / 3	2 / 2
0006397 mRNA processing	232	1 / 4	0 / 3	3 / 3	1 / 3	1 / 2	3 / 2	8 / 22	**61 / 33**	16 / 16	8 / 12
0051169 nuclear transport	120	3 / 2	2 / 2	0 / 1	5 / 2	0 / 1	0 / 1	6 / 12	**38 / 17**	11 / 8	9 / 6
0006406 mRNA export from nucleus	36	0 / 1	0 / 1	0 / 0	0 / 1	0 / 0	0 / 0	1 / 4	**17 / 5**	2 / 2	2 / 2
0007049 cell cycle	737	9 / 13	6 / 10	10 / 9	11 / 10	7 / 7	4 / 8	74 / 71	**148 / 104**	**73 / 49**	30 / 39
0007126 meiosis	65	2 / 1	1 / 1	0 / 1	2 / 1	0 / 1	0 / 1	3 / 6	10 / 9	**20 / 4**	6 / 3
0007286 spermatid development	30	0 / 1	0 / 0	0 / 0	0 / 0	0 / 0	1 / 0	0 / 3	2 / 4	**10 / 2**	**9 / 2**
0009566 fertilization	39	1 / 1	0 / 1	1 / 1	1 / 1	0 / 0	0 / 0	2 / 4	1 / 6	**10 / 3**	**12 / 2**
0007283 spermatogenesis	179	1 / 3	1 / 3	0 / 2	1 / 2	1 / 2	3 / 2	7 / 17	15 / 25	**47 / 12**	**41 / 10**
0000003 reproduction	335	6 / 6	1 / 5	4 / 4	2 / 4	1 / 3	3 / 4	22 / 32	45 / 47	**60 / 22**	**70 / 18**
0007001 chromosome organization and biogenesis	316	8 / 5	2 / 4	2 / 4	9 / 4	3 / 3	2 / 3	**7 / 30**	**65 / 45**	34 / 21	**36 / 17**
0007342 fusion of sperm to egg plasma membrane	9	1 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	1 / 1	0 / 1	0 / 1	**6 / 1**

Figure 99. Biological processes enriched in the selected groups of co-expressed genes.
Each cluster is matched with enriched GO terms from the ontology "biological process". For each group, the number of genes annotated with a biological process is indicated below the group name. The number of genes associated with a particular biological process GO term is indicated to the right of this annotation. Numbers of genes associated with a specific GO term and enriched in each cluster are given within rectangles in bold as observed and as expected. A color code indicates enrichment (red) or depletion (blue) as indicated in the scale bar in the top left corner.

As expected, this analysis demonstrates that the gene groups G2 and G6 with an expression profile similar to rod photoreceptor marker genes are significantly enriched in vision processes (*visual perception and phototransduction*).

This GO analysis demonstrates that both ME and PM gene groups, showing a peak of expression in spermatocytes and spermatids respectively, are significantly enriched in *reproduction* and *spermatogenesis.* Nevertheless, the ME group is specialized in *meiosis* while the PM group is specialized in the final germ cell differentiation. We will thus further focus our analysis on the PM cluster.

As a second step of filtration, genes specifically expressed in retina or in testis tissues were further selected. Thus, a gene is selected as retinal specific if its unlog expression value is >100 in retina and <100 in all other tissues (excluding the eye and brain samples), with a minimal fold-change of 2 between the retina sample and the other samples. Similarly, a gene is selected as testis specific if its unlog expression value is >100 in germ cells and <100 in all other tissues with a minimal fold-change of 2 between the germ cells and the other samples.

The heatmap of the genes selected as retina- or testis-specific in each group of co-expressed genes is shown in Figure 100.

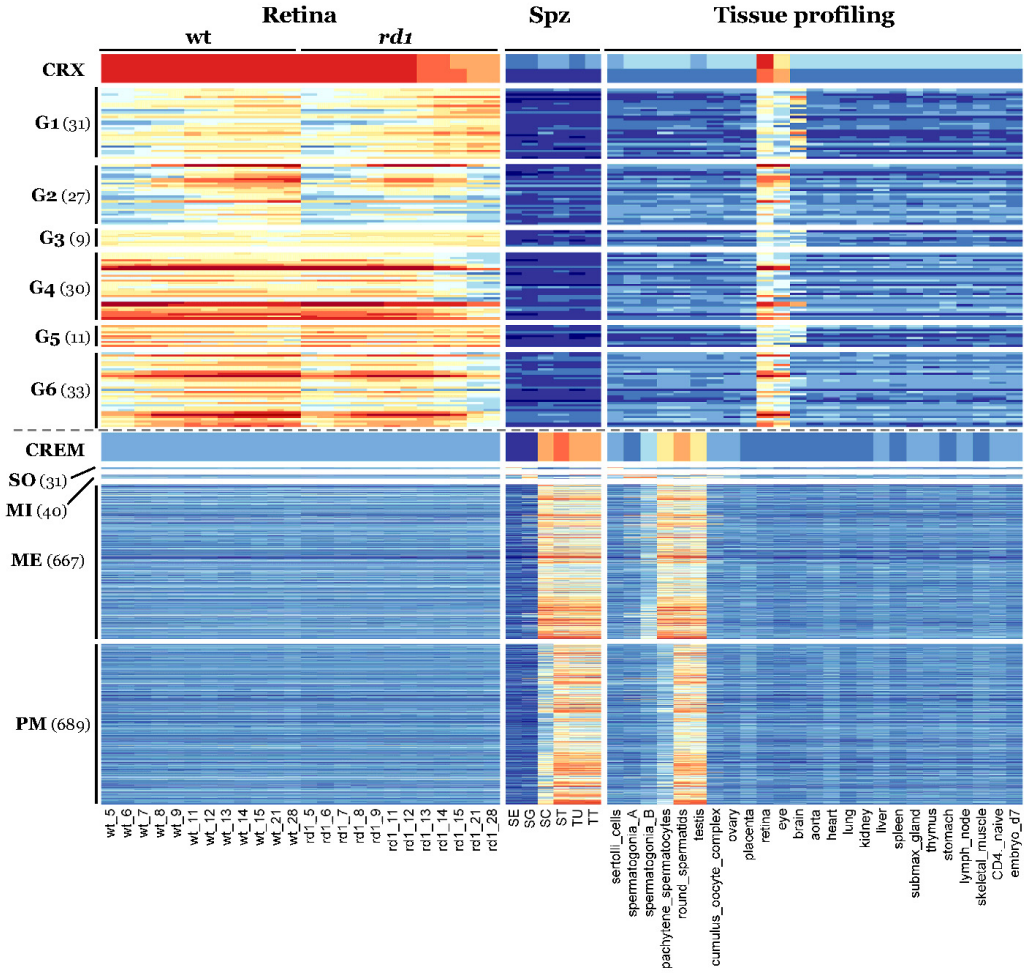


Figure 100. Heatmap of the selected groups of retina-specific genes co-expressed in the *rd1* and wt retina (noted as Retina in the upper part), of the selected groups of testis-specific genes co-expressed in germ cells (Spz), and of tissue profiling transcriptomics data (Tissue profiling).
The expression data of the transcripts selected as retina specific in the six groups determined from the *rd1* transcriptome analysis are shown above the grey dashed line. Similarly, the expression data of the transcripts selected as testis specific in the four groups of the mouse spermatogenesis study are shown below the grey dashed line. The expression data of the CRX and CREM TFs are also shown for all different conditions. The color code is the same as in Figure 98.

The heatmap illustrates the tissue specificity of the selected genes. It should be noted that the SO and MI clusters are not testis-specific. They both contain about 1% of testis-specific genes.

Because of calculation time, we focus our analysis of TFBS enrichment on two well studied TFs: the cone-rod homeobox (CRX) and c-AMP Response Element Modulator (CREM). Indeed, CRX and CREM play an important role in the regulation of retina and testis genes respectively (Qian et al., 2005) (Hogeveen and Sassone-Corsi, 2006) and are themselves

specifically expressed in retina and in testis (see Figure 100). CRX ans CREM perfectly fit the expression profiles of the genes belonging to the selected *rd1* clusters (G2 and G6) and the selected spermatogenesis cluster (PM) (Figure 100).

In such stringent conditions of co-expression we hypothesise that these remaining genes (preferentially expressed in a given cell types of a tissue and nowhere else in the body) are co-regulated. Therefore, we investigated the possibility of a potential enrichment in these clusters of photoreceptor-specific (CRX) and post-meiotic specific (CREM) TFs.

13.2 CRX and CREM TFBS enrichment calculation

13.2.1 CRX motif

The CRX motifs were predicted using HT PromAn for two sets of genomic sequences corresponding to 1,000 randomly selected genes, as well as for the G2, G6 and PM tissue-specific clusters (see section 6.3). The consensus sequence YTAATCM of the CRX TFBS was considered (Pittler et al., 2004).

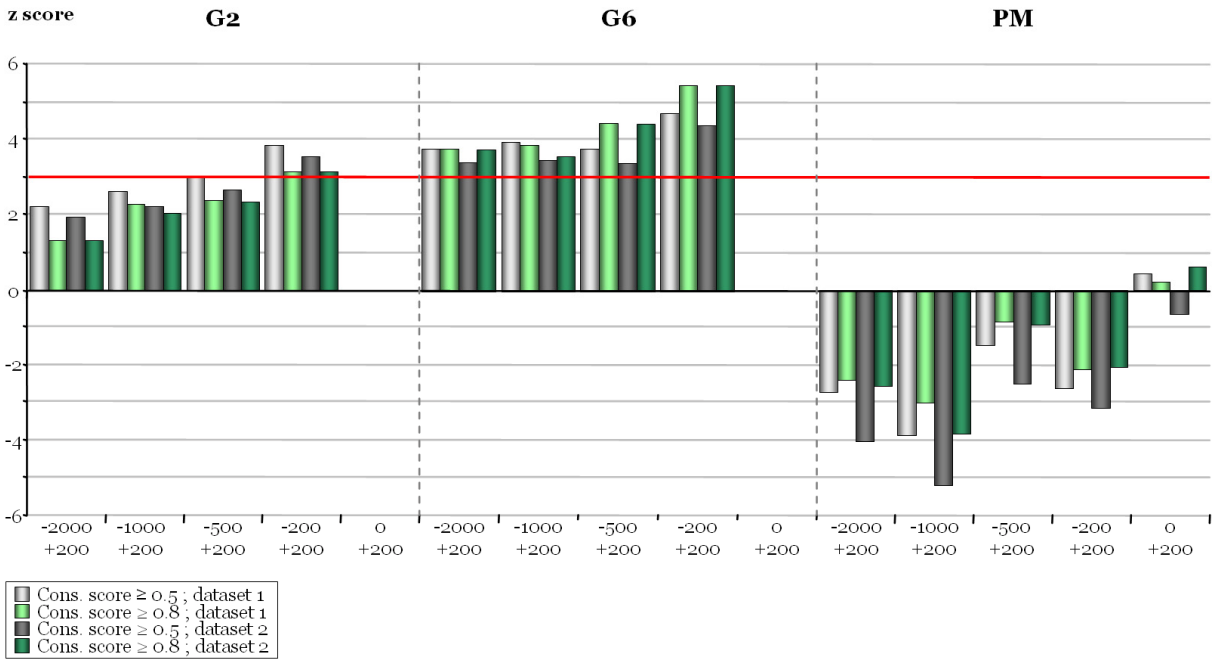


Figure 101. CRX TFBS enrichments in the G2, G6 and PM tissue-specific clusters.
The CRX motif enrichments were calculated in several regions (-2,000/+200, -1,000/+200, -500/+200, -200/+200, 0/+200) relative to the TSS position, with a matrix score >0.8 and a conservation score >0.5 (in grey) or >0.8 (in green). The enrichment calculations were performed by comparison to two datasets of 1,000 randomly selected sequences; *dataset1* is shown with light colors and *dataset2* with dark colors. At least three motifs must be found in each group to calculate the over-representation (see section Chapter 8 -). The enrichments were estimated with *z score* values, where a positive *z score* value indicates an enrichment. Enrichment values >3 (probability of 0.00135, red line) are considered to be highly significant.

The cluster G2 is slightly enriched (*z score* >3, shown as a red line) in the proximal region from -200 to +200 relative to the TSS, with conservation scores of either 0.5 or 0.8. Among the genes belonging to the G2 cluster and having a CRX TFBS automatically detected in their promoter region, we found the Rp1h (retinitis pigmentosa 1 homolog) and the Pde6a (phosphodiesterase 6A) genes that are known to be directly regulated by CRX. Indeed, three experimental assays highly suggest that RP1 is indeed *bona fide* targets of CRX in vivo (Qian et al., 2005). Pde6a mRNA is reduced by 87% in the retina of CRX(-/-) mice and is undetectable in NRL(-/-) mice at postnatal day 10 that is consistent with a requirement for CRX and NRL in Pde6a promoter activity (Pittler et al., 2004).

In G6, CRX TFBSs are over-represented in all the regions upstream of the TSS, with a peak of enrichment in the region proximal to the TSS. Among the genes belonging to the G6 cluster and having a CRX TFBS automatically detected in their promoter region, we found the Rho (rhodopsin) and Rbp3 (Retinol binding protein 3 or IRBP) genes that are known to be directly regulated by CRX. In transient transfection studies, it has been reported that CRX

transactivates rhodopsin promoter-reporter constructs (Chen et al., 1997). Moreover NRL and CRX act synergistically to regulate rhodopsin transcription (Mears et al., 2001). It is also well known that the CRX interaction with CRXE is essential for the photoreceptor-specific activity of the IRBP promoter in vivo (Fei et al., 1999).

In G2 and G6, CRX TFBSs are absent from the region immediately downstream of the TSS (0/+200). In contrast, as might be expected, the PM cluster shows no enrichment in the CRX TFBS and in fact, an under-representation of CRX TFBSs is observed in regions downstream of the TSS.

13.2.2 CREM motif

The CREM TF recognizes CRE (c-AMP Response Element) motifs which were predicted for the same two sets of genomic sequences corresponding to 1,000 random selected genes as well as on the G2, G6 and PM tissue-specific clusters (see section 6.3). The consensus sequence TGACGTCA of the CRE TFBS was considered (Lui et al., 2006).

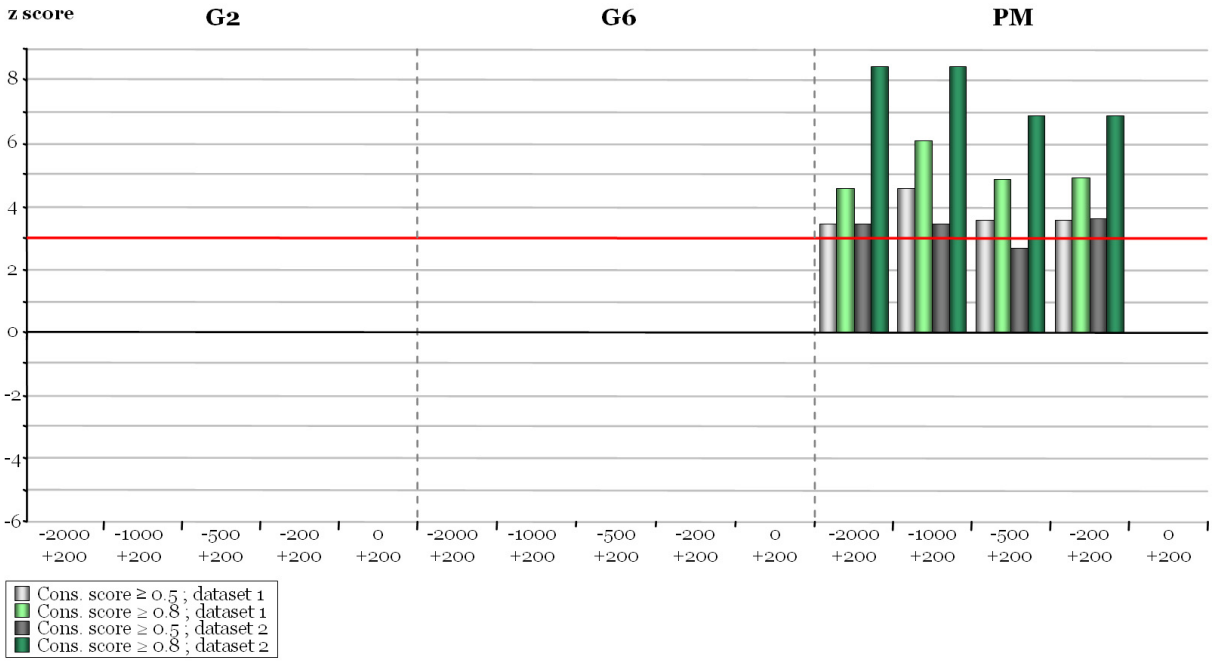


Figure 102. CREM enrichments in the G2, G6 and PM tissue-specific clusters.
See legend of the Figure 101.

The CREM TFBS is totally absent from the G2 and G6 clusters. As expected, the PM cluster shows a significant enrichment in CREM TFBS in all the regions downstream of the TSS. Among the genes belonging to the PM cluster and having a CREM TFBS automatically detected in their promoter region, we found the KIF17b gene. The KIF17b protein is involved

in nucleocytoplasmic transport of RNA and of transcriptional coactivator (Kotaja et al., 2006). Strikingly, the motor domain NOR on microtubules of KIF17b allow this protein to shuttle between nuclear and cytoplasmic compartments and to transport activator of CREM in testis such as ACT (Fimia et al., 1999) (Kotaja et al., 2005). The ACT-KIF17b interaction is restricted to specific stages of spermatogenesis and directly determines the intracellular localization of ACT (Macho et al., 2002). These results suggest that CREM regulates the transcription of the KIF17b gene involved in the nuclear transport of the ACT testis-specific coactivator which regulates the transcription of CREM during spermatogenesis. In other words, CREM would indirectly regulate its own expression in testis.

As previously observed for CRX in the G2 and G6 clustersn, the CREM TFBS is also absent from the regions immediately downstream of the TSS (0/+200).

13.3 Discussion and perspectives

These analyses performed on highly specialised biological systems with a stringent selection of differentially expressed genes, which are further clustered into groups of co-expressed and tissue-specific genes, allowed us to identify potential co-regulation through high-throughput enrichment studies of specific TFs. We observed that clusters of genes preferentially expressed in photoreceptors (G2 and G6) and deregulated during the *rd1* degeneration process are enriched in photoreceptor-specific CRX TFBSs whereas this motif is under-represented in the PM cluster. Similarly testis-specific genes with a peak of expression in spermatids (PM) were enriched in the CREM TFBS, which is involved in post-meiotic transcriptional regulation. This TFBS is totally absent from retina-specific clusters. It is important to note that the TFBS search was performed using very stringent parameters: matrix score >0.8 and conservation score estimated on an alignment of genomic sequences from 17 vertebrates >0.5 or >0.8. Such conditions allowed us to show significant TFBS enrichment in clusters having similar expression profiles as the considered TF, but also an under-representation of these TFBSs in other biological systems using exactly the same parameters. These results clearly illustrate the importance of an integrative strategy combining very distinct complementary high throughput approaches for efficient promoter analysis. In the presented studies, these approaches encompass not only the use of successful predictive promoter tools with different algorithms and databases but also the definition of valuable gene functions through the use of GO term enrichment statistics as well as the analysis and cross-validation of complete transcriptomics data sets. One problem common to all the data processed, concerns the notion of enrichment and notably the estimation of its statistical significance. Indeed, the high throughput data currently being generated in

functional genomics and biology are frequently partial, poorly reproductible, noisy and redundant. All these features represent major challenges for statistical analyses that can be overcome by robust *in silico* analysis combined with systematic experimental validation which will not only validate the bioinformatics predictions but also will allow a sub-sequent refinement of the analysis strategy.

In this context, in addition to the experimental validation, it would be very interesting to complement our analysis with other retina-specific TFs such as *nrl* and *nr2e3* (Qian et al., 2005) and with other testis-specific TFs such as *Tcfl5* (Siep et al., 2004) to verify that our approach is effective. Moreover, the enrichments were calculated using two random datasets of 1,000 genes but it would be statistically more relevant to calculate the enrichment in comparison to the complete genome. Indeed, in the case of the CREM TFBSs in a PM cluster with a conservation score > 0.8, we observe an enrichment with both random datasets but this is much higher in the dataset 2 (e.g. between -200 and +200, *z score* =7 in the dataset 1 compared to 5 in the dataset 2). Furthermore, while the CRX TFBS seems to be preferentially located in the proximal region of the promoter (up to -500 bp), the CREM TFBS does not seem to have a preference of localization relative to the TSS. Verification of this observation would require several weeks of intensive calculation and a significant amount of expert interpretation. It would also require a prior-optimization of conditions allowing a particularly stringent selection of the analysed sets of genes. Nevertheless, a balance has to be found because overly stringent parameters would reduce the 25,000 known human genes to only a handful of genes which would not permit significant enrichment calculations. The aim is to minimize the noise without losing the entire signal.

It would be also very interesting to verify whether such an integrative strategy applied on several TFBSs with combinatorial statistics would allow one to highlight combination of different TFBSs (CRMs) responsible for tissue-specificity or specific biological processes.

Conclusions and perspectives

The post-genomic era has contributed to our knowledge in many scientific fields thanks to the availability of complete genome sequences and of high-throughput technologies. One of the most revolutionized fields is probably the gene concept, both at the level of the gene transcriptional products and at the level of the regulation of gene expression. The former is linked to the understanding of the incredible complexity of gene codage, while the latter concerns the new opportunities introduced by in silico promoter analysis which now complement the experimental approaches.

Initially, bioinformatics studies of the regulation of gene expression were performed using the classical approach to prediction algorithm development, starting with a primary collection and analysis of a training set (here promoter elements), followed by the definition of rules that might be suitable for efficient prediction. However, this strategy failed as the signal in promoter sequences corresponds to small, highly variable sequences composed of four letters (nucleotides) named Transcription Factor Binding Sites (TFBSs). The bona fide TFBSs constitute a very weak signal among an extremely noisy environment of predictions. In addition, whereas the start of translation is defined by the start codon, no rule exists for the start of transcription. Thus, the current algorithmic methods that exist for the prediction of promoter sequences and Transcriptional Start Sites (TSSs) localization are still unreliable and data produced by bioinformatics predictions include a huge amount of false positives that needs to be minimized. One possible solution to this problem is to identify a "sheltered environment" in which the specificity of pattern discovery might be enhanced. For example, evolutionary information can be exploited, based on the assumption that regulatory elements in non-coding regions are under a higher selective pressure during evolution than non-functional genomic regions.

In this context, PromAn provides an innovative integrative approach in the form of a versatile, modular and flexible program dedicated to the analysis of promoter regions. The program was implemented as an intuitive tool allowing expert-guided as well as automatic analysis of promoter regions. PromAn integrates several approaches aimed at the localization of TSSs and promoter regions and the detection of biologically active TFBSs through efficient filtering. As shown in the manuscript, the PromAn program has been successfully exploited in a number of biological projects which each represent a distinct challenge for promoter prediction, ranging from the refinement of the location of a promoter or the relevant prediction of bona fide TFBSs in a gene family, up to the high throughput analysis of thousands of genes showing common transcriptomic expression patterns. In all cases, the rationale was to perform the bioinformatics prediction on large genomic regions in combination with experimental validations. Indeed, the development of an integrative

prediction program and experimental analysis can no longer be considered as separate problems, because in silico and experimental results must be considered together in order to guide and cross-validate the global strategy.

The study of the promoter sequence on a large genomic region has proved to be important to visualize the global genomic context of the gene. Indeed, the orientation of neighbouring genes separated by small intergenic regions, the presence of regulatory elements in introns, untranslated regions or distal regions upstream of the TSS have a great effect on the expression regulation of the gene. In the future, this strategy will allow a better integration of important transcriptional features such as potentially active bidirectional promoters, regulatory elements downstream of the TSS or regulation by non-coding RNA. Similarly, distal regulatory elements upstream of the TSS will be taken into account as they can be implicated in a "looping" mechanism of the DNA to increase the density of TFBSs in the vicinity of the proximal promoter.

Furthermore, tissue-specific expression patterns of the transcription factors (TFs) allowed a subsequent filtering of the TFBS predictions for given experimental conditions. Indeed, genes with a tissue-specific expression should be regulated by TFs expressed in these tissues. This highlights that data describing direct protein-protein physical interactions between transcription factors and/or co-activators will greatly improve the identification of cis-regulatory modules (combinations of close TFBSs), and increase our understanding of the global mechanisms yielding to the expression regulation of a single gene. Taken together, these complementary approaches applied to a population of co-expressed genes in a given biological system will shed light on the elaboration of complex regulatory networks with important implications in human health and disease.

Ultimately, the accumulated knowledge of gene expression regulation mechanisms will lead to the understanding of new levels of integration playing a critical role in the final gene and genome codage power. These higher levels are characterized by the complexity and the multiplicity of new regulatory actors such as the state of compaction of the chromatin, the temporal stages and the cellular gene environment.

References

Abou-Sleymane, G., Chalmel, F., Helmlinger, D., Lardenois, A., Thibault, C., Weber, C., Merienne, K., Mandel, J. L., Poch, O., Devys, D., and Trottier, Y. (2006). Polyglutamine expansion causes neurodegeneration by altering the neuronal differentiation program. Hum Mol Genet *15*, 691-703.

Adachi, N., and Lieber, M. R. (2002). Bidirectional gene organization: a common architectural feature of the human genome. Cell *109*, 807-809.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., and et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science *252*, 1651-1656.

Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y., and De Moor, B. (2003). Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Res *31*, 1753-1764.

Aerts, S., Thijs, G., Dabrowski, M., Moreau, Y., and De Moor, B. (2004). Comprehensive analysis of the base composition around the transcription start site in Metazoa. BMC Genomics *5*, 34.

Ahn, J., and Gruen, J. R. (1999). The genomic organization of the histone clusters on human 6p21.3. Mamm Genome *10*, 768-770.

Albig, W., Kioschis, P., Poustka, A., Meergans, K., and Doenecke, D. (1997). Human histone gene organization: nonregular arrangement within a large cluster. Genomics *40*, 314-322.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

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.

Anfinsen, C. B., and Redfield, R. R. (1956). Protein structure in relation to function and biosynthesis. Adv Protein Chem *48*, 1-100.

Antequera, F., and Bird, A. (1993). Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A *90*, 11995-11999.

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.

Arnone, M. I., and Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. Development *124*, 1851-1864.

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.

Audit, B., Vaillant, C., Arneodo, A., d'Aubenton-Carafa, Y., and Thermes, C. (2002). Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes. J Mol Biol *316*, 903-918.

Bagga, R., Michalowski, S., Sabnis, R., Griffith, J. D., and Emerson, B. M. (2000). HMG I/Y regulates long-range enhancer-dependent transcription on DNA and chromatin by changes in DNA topology. Nucleic Acids Res *28*, 2541-2550.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., *et al.* (2005). The Universal Protein Resource (UniProt). Nucleic Acids Res *33*, D154-159.

Bajic, V. B., Choudhary, V., and Hock, C. K. (2004a). Content analysis of the core promoter region of human genes. In Silico Biol *4*, 109-125.

Bajic, V. B., and Seah, S. H. (2003). Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units. Genome Res *13*, 1923-1929.

Bajic, V. B., Seah, S. H., Chong, A., Krishnan, S. P., Koh, J. L., and Brusic, V. (2003). Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates. J Mol Graph Model *21*, 323-332.

Bajic, V. B., Seah, S. H., Chong, A., Zhang, G., Koh, J. L., and Brusic, V. (2002). Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. Bioinformatics *18*, 198-199.

Bajic, V. B., Tan, S. L., Christoffels, A., Schonbach, C., Lipovich, L., Yang, L., Hofmann, O., Kruger, A., Hide, W., Kai, C., *et al.* (2006). Mice and men: their promoter properties. PLoS Genet *2*, e54.

Bajic, V. B., Tan, S. L., Suzuki, Y., and Sugano, S. (2004b). Promoter prediction analysis on the whole human genome. Nat Biotechnol *22*, 1467-1473.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Barton, L. M., Gottgens, 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.

Barton, M. C., Madani, N., and Emerson, B. M. (1997). Distal enhancer regulation by promoter derepression in topologically constrained DNA in vitro. Proc Natl Acad Sci U S A *94*, 7257-7262.

Basehoar, A. D., Zanton, S. J., and Pugh, B. F. (2004). Identification and distinct regulation of yeast TATA box-containing genes. Cell *116*, 699-709.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321-1325.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2005). GenBank. Nucleic Acids Res *33*, D34-38.

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.

Bird, A. P., and Taggart, M. H. (1980). Variable patterns of total DNA and rDNA methylation in animals. Nucleic Acids Res *8*, 1485-1497.

Blader, P., Plessy, C., and Strahle, U. (2003). Multiple regulatory elements with spatially and temporally distinct activities control neurogenin1 expression in primary neurons of the zebrafish embryo. Mech Dev *120*, 211-218.

Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganiere, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D., *et al.* (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res *16*, 656-668.

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.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., *et al.* (1997). The complete genome sequence of Escherichia coli K-12. Science *277*, 1453-1474.

Bode, J., Benham, C., Knopp, A., and Mielke, C. (2000). Transcriptional augmentation: modulation of gene expression by scaffold/matrix-attached regions (S/MAR elements). Crit Rev Eukaryot Gene Expr *10*, 73-90.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., *et al.* (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res *31*, 365-370.

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.

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993). dbEST--database for "expressed sequence tags". Nat Genet *4*, 332-333.

Bortoluzzi, S., Coppe, A., Bisognin, A., Pizzi, C., and Danieli, G. A. (2005). A multistep bioinformatic approach detects putative regulatory elements in gene promoters. BMC Bioinformatics *6*, 121.

Borukhov, S., and Nudler, E. (2003). RNA polymerase holoenzyme: structure, function and biological implications. Curr Opin Microbiol *6*, 93-100.

Brahmachari, S. K., Meera, G., Sarkar, P. S., Balagurumoorthy, P., Tripathi, J., Raghavan, S., Shaligram, U., and Pataskar, S. (1995). Simple repetitive sequences in the genome: structure and functional significance. Electrophoresis *16*, 1705-1714.

Bray, N., Dubchak, I., and Pachter, L. (2003). AVID: A global alignment program. Genome Res *13*, 97-102.

Brosius, J. (1999). RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene *238*, 115-134.

Brown, R. P., and Feder, M. E. (2005). Reverse transcriptional profiling: non-correspondence of transcript level variation and proximal promoter polymorphism. BMC Genomics *6*, 110.

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.

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.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol *212*, 563-578.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol *268*, 78-94.

Burgess-Beusse, B., Farrell, C., Gaszner, M., Litt, M., Mutskov, V., Recillas-Targa, F., Simpson, M., West, A., and Felsenfeld, G. (2002). The insulation of genes from external enhancers and silencing chromatin. Proc Natl Acad Sci U S A *99 Suppl 4*, 16433-16437.

Burke, D. T., Carle, G. F., and Olson, M. V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science *236*, 806-812.

Burke, T. W., and Kadonaga, J. T. (1996). Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev *10*, 711-724.

Burke, T. W., Willy, P. J., Kutach, A. K., Butler, J. E., and Kadonaga, J. T. (1998). The DPE, a conserved downstream core promoter element that is functionally analogous to the TATA box. Cold Spring Harb Symp Quant Biol *63*, 75-82.

Bussemaker, H. J., Li, H., and Siggia, E. D. (2001). Regulatory element detection using correlation with expression. Nat Genet *27*, 167-171.

Butler, B. A. (1998). Sequence analysis using GCG. Methods Biochem Anal *39*, 74-97.

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

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.

Camarero, N., Mascaro, C., Mayordomo, C., Vilardell, F., Haro, D., and Marrero, P. F. (2006). Ketogenic HMGCS2 Is a c-Myc target gene expressed in differentiated cells of human colonic epithelium and down-regulated in colon cancer. Mol Cancer Res *4*, 645-653.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., *et al.* (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559-1563.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstrom, P. G., Frith, M. C., *et al.* (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet *38*, 626-635.

Cavener, D. R. (1987). Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. Nucleic Acids Res *15*, 1353-1361.

Cavin Perier, R., Junier, T., and Bucher, P. (1998). The Eukaryotic Promoter Database EPD. Nucleic Acids Res *26*, 353-357.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., *et al.* (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell *116*, 499-509.

Chalmel, F., Lardenois, A., Thompson, J. D., Muller, J., Sahel, J. A., Leveillard, T., and Poch, O. (2005). GOAnno: GO annotation based on multiple alignment. Bioinformatics *21*, 2095-2096.

Chang, B., Heckenlively, J. R., Hawes, N. L., and Roderick, T. H. (1993). New mouse primary retinal degeneration (rd-3). Genomics *16*, 45-49.

Chang, L. W., Nagarajan, R., Magee, J. A., Milbrandt, J., and Stormo, G. D. (2006). A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. Genome Res *16*, 405-413.

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.

Chen, S., Wang, Q. L., Nie, Z., Sun, H., Lennon, G., Copeland, N. G., Gilbert, D. J., Jenkins, N. A., and Zack, D. J. (1997). Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. Neuron *19*, 1017-1030.

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.

Cho, R. J., Huang, M., Campbell, M. J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S. J., Davis, R. W., and Lockhart, D. J. (2001). Transcriptional regulation and function during the human cell cycle. Nat Genet *27*, 48-54.

Chou, P. Y., and Fasman, G. D. (1974). Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry *13*, 211-222.

Christoffels, V. M., Grange, T., Kaestner, K. H., Cole, T. J., Darlington, G. J., Croniger, C. M., and Lamers, W. H. (1998). Glucocorticoid receptor, C/EBP, HNF3, and protein kinase A coordinately activate the glucocorticoid response unit of the carbamoylphosphate synthetase I gene. Mol Cell Biol *18*, 6305-6315.

Christy, R. J., Yang, V. W., Ntambi, J. M., Geiman, D. E., Landschulz, W. H., Friedman, A. D., Nakabeppu, Y., Kelly, T. J., and Lane, M. D. (1989). Differentiation-induced gene expression in 3T3-L1

preadipocytes: CCAAT/enhancer binding protein interacts with and activates the promoters of two adipocyte-specific genes. Genes Dev *3*, 1323-1335.

Cohen, D., Chumakov, I., and Weissenbach, J. (1993). A first-generation physical map of the human genome. Nature *366*, 698-701.

Collins, F. S. (1990). Identifying human disease genes by positional cloning. Harvey Lect *86*, 149-164.

Compe, E., Drane, P., Laurent, C., Diderich, K., Braun, C., Hoeijmakers, J. H., and Egly, J. M. (2005). Dysregulation of the peroxisome proliferator-activated receptor target genes by XPD mutations. Mol Cell Biol *25*, 6065-6076.

Cooper, G. M., and Sidow, A. (2003). Genomic regulatory regions: insights from comparative sequence analysis. Curr Opin Genet Dev *13*, 604-610.

Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L., and Myers, R. M. (2006). Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res *16*, 1-10.

Cora, D., Herrmann, C., Dieterich, C., Di Cunto, F., Provero, P., and Caselle, M. (2005). Ab initio identification of putative human transcription factor binding sites by comparative genomics. BMC Bioinformatics *6*, 110.

Corcoran, D. L., Feingold, E., and Benos, P. V. (2005). FOOTER: a web tool for finding mammalian DNA regulatory regions using phylogenetic footprinting. Nucleic Acids Res *33*, W442-446.

Costello, J. F., Fruhwald, M. C., Smiraglia, D. J., Rush, L. J., Robertson, G. P., Gao, X., Wright, F. A., Feramisco, J. D., Peltomaki, P., Lang, J. C., *et al.* (2000). Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. Nat Genet *24*, 132-138.

Crick, F. H. (1958). On protein synthesis. Symp Soc Exp Biol *12*, 138-163.

Cross, S. H., and Bird, A. P. (1995). CpG islands and genes. Curr Opin Genet Dev *5*, 309-314.

Cullen, B. R. (2002). RNA interference: antiviral defense and genetic tool. Nat Immunol *3*, 597-599.

Davuluri, R. V., Grosse, I., and Zhang, M. Q. (2001). Computational identification of promoters and first exons in the human genome. Nat Genet *29*, 412-417.

Dayhoff, M. O. (1965). Computer aids to protein sequence determination. J Theor Biol *8*, 97-112.

de Laat, W., and Grosveld, F. (2003). Spatial organization of gene expression: the active chromatin hub. Chromosome Res *11*, 447-459.

Deng, W., and Roberts, S. G. (2005). A core promoter element downstream of the TATA box that is recognized by TFIIB. Genes Dev *19*, 2418-2423.

Dennis, C., and Surridge, C. (2000). Arabidopsis thaliana genome. Introduction. Nature *408*, 791.

Dermitzakis, E. T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B. J., Flegel, V., Bucher, P., Jongeneel, C. V., and Antonarakis, S. E. (2002). Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature *420*, 578-582.

Dermitzakis, E. T., Reymond, A., Scamuffa, N., Ucla, C., Kirkness, E., Rossier, C., and Antonarakis, S. E. (2003). Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). Science *302*, 1033-1035.

Di Cara, A., Schmidt, K., Hemmings, B. A., and Oakeley, E. J. (2005). PromoterPlot: a graphical display of promoter similarities by pattern recognition. Nucleic Acids Res *33*, W423-426.

Dickmeis, T., Plessy, C., Rastegar, S., Aanstad, P., Herwig, R., Chalmel, F., Fischer, N., and Strahle, U. (2004). Expression profiling and comparative genomics identify a conserved regulatory region controlling midline expression in the zebrafish embryo. Genome Res *14*, 228-238.

Distel, R. J., Ro, H. S., Rosen, B. S., Groves, D. L., and Spiegelman, B. M. (1987). Nucleoprotein complexes that regulate gene expression in adipocyte differentiation: direct participation of c-fos. Cell *49*, 835-844.

Donaldson, I. J., Chapman, M., Kinston, S., Landry, J. R., Knezevic, K., Piltz, S., Buckley, N., Green, A. R., and Gottgens, B. (2005). Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. Hum Mol Genet *14*, 595-601.

Dong, S., Lester, L., and Johnson, L. F. (2000). Transcriptional control elements and complex initiation pattern of the TATA-less bidirectional human thymidylate synthase promoter. J Cell Biochem *77*, 50-64.

Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol *4*, R7.

Down, T. A., and Hubbard, T. J. (2002). Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res *12*, 458-461.

Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., *et al.* (1999). The DNA sequence of human chromosome 22. Nature *402*, 489-495.

Eisenberg, E., and Levanon, E. Y. (2003). Human housekeeping genes are compact. Trends Genet *19*, 362-365.

Elkon, R., Linhart, C., Sharan, R., Shamir, R., and Shiloh, Y. (2003). Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. Genome Res *13*, 773-780.

Elnitski, L., Giardine, B., Shah, P., Zhang, Y., Riemer, C., Weirauch, M., Burhans, R., Miller, W., and Hardison, R. C. (2005). Improvements to GALA and dbERGE II: databases featuring genomic sequence alignment, annotation and experimental results. Nucleic Acids Res *33*, D466-470.

Elnitski, L., Hardison, R. C., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M. J., Schwartz, S., Miller, W., and Chiaromonte, F. (2003). Distinguishing regulatory DNA from neutral sites. Genome Res *13*, 64-72.

Engstrom, P. G., Suzuki, H., Ninomiya, N., Akalin, A., Sessa, L., Lavorgna, G., Brozzi, A., Luzi, L., Tan, S. L., Yang, L., *et al.* (2006). Complex Loci in human and mouse genomes. PLoS Genet *2*, e47.

Etzold, T., and Argos, P. (1993). SRS--an indexing and retrieval tool for flat file data libraries. Comput Appl Biosci *9*, 49-57.

Fei, Y., Matragoon, S., Smith, S. B., Overbeek, P. A., Chen, S., Zack, D. J., and Liou, G. I. (1999). Functional dissection of the promoter of the interphotoreceptor retinoid-binding protein gene: the cone-rod-homeobox element is essential for photoreceptor-specific expression in vivo. J Biochem (Tokyo) *125*, 1189-1199.

Fickett, J. W., and Wasserman, W. W. (2000). Discovery and modeling of transcriptional regulatory regions. Curr Opin Biotechnol *11*, 19-24.

Fimia, G. M., De Cesare, D., and Sassone-Corsi, P. (1999). CBP-independent activation of CREM and CREB by the LIM-only protein ACT. Nature *398*, 165-169.

Fitch, W. M., and Margoliash, E. (1967). Construction of phylogenetic trees. Science *155*, 279-284.

Fitzgerald, P. C., Sturgill, D., Shyakhtenko, A., Oliver, B., and Vinson, C. (2006). Comparative genomics of Drosophila and human core promoters. Genome Biol *7*, R53.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., and et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496-512.

Florquin, K., Saeys, Y., Degroeve, S., Rouze, P., and Van de Peer, Y. (2005). Large-scale structural analysis of the core promoter in mammalian and plant genomes. Nucleic Acids Res *33*, 4255-4264.

Frazer, K. A., Elnitski, L., Church, D. M., Dubchak, I., and Hardison, R. C. (2003). Cross-species sequence comparisons: a review of methods and available resources. Genome Res *13*, 1-12.

Frickey, T., and Lupas, A. N. (2004). PhyloGenie: automated phylome generation and analysis. Nucleic Acids Res *32*, 5231-5238.

Frith, M. C., Ponjavic, J., Fredman, D., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Sandelin, A. (2006). Evolutionary turnover of mammalian transcription start sites. Genome Res *16*, 713-722.

Fukue, Y., Sumida, N., Nishikawa, J., and Ohyama, T. (2004). Core promoter elements of eukaryotic genes have a highly distinctive mechanical property. Nucleic Acids Res *32*, 5834-5840.

Ganapathi, M., Srivastava, P., Das Sutar, S. K., Kumar, K., Dasgupta, D., Pal Singh, G., Brahmachari, V., and Brahmachari, S. K. (2005). Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes. BMC Bioinformatics *6*, 126.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J Mol Biol *196*, 261-282.

Gibbs, R. A., and Nelson, D. L. (2003). Human genetics. Primate shadow play. Science *299*, 1331-1333.

Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E., *et al.* (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature *428*, 493-521.

Glazko, G. V., Koonin, E. V., Rogozin, I. B., and Shabalina, S. A. (2003). A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. Trends Genet *19*, 119-124.

Glazko, G. V., Rogozin, I. B., and Glazkov, M. V. (2001). Comparative study and prediction of DNA fragments associated with various elements of the nuclear matrix. Biochim Biophys Acta *1517*, 351-364.

Goncalves, I., Duret, L., and Mouchiroud, D. (2000). Nature and structure of human genes that generate retropseudogenes. Genome Res *10*, 672-678.

Goode, D. K., Snell, P., Smith, S. F., Cooke, J. E., and Elgar, G. (2005). Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. Genomics *86*, 172-181.

Gouy, M., Milleret, F., Mugnier, C., Jacobzone, M., and Gautier, C. (1984). ACNUC: a nucleic acid sequence data base and analysis system. Nucleic Acids Res *12*, 121-127.

Graves, R. A., Tontonoz, P., Platt, K. A., Ross, S. R., and Spiegelman, B. M. (1992). Identification of a fat cell enhancer: analysis of requirements for adipose tissue-specific gene expression. J Cell Biochem *49*, 219-224.

Grewal, S. I., and Moazed, D. (2003). Heterochromatin and epigenetic control of gene expression. Science *301*, 798-802.

Grover, D., Mukerji, M., Bhatnagar, P., Kannan, K., and Brahmachari, S. K. (2004). Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition. Bioinformatics *20*, 813-817.

Guarguaglini, G., Battistoni, A., Pittoggi, C., Di Matteo, G., Di Fiore, B., and Lavia, P. (1997). Expression of the murine RanBP1 and Htf9-c genes is regulated from a shared bidirectional promoter during cell cycle progression. Biochem J *325 (Pt 1)*, 277-286.

GuhaThakurta, D., Palomar, L., Stormo, G. D., Tedesco, P., Johnson, T. E., Walker, D. W., Lithgow, G., Kim, S., and Link, C. D. (2002). Identification of a novel cis-regulatory element involved in the heat shock response in Caenorhabditis elegans using microarray gene expression and computational methods. Genome Res *12*, 701-712.

Gustincich, S., Sandelin, A., Plessy, C., Katayama, S., Simone, R., Lazarevic, D., Hayashizaki, Y., and Carninci, P. (2006). The complexity of the mammalian transcriptome. J Physiol *575*, 321-332.

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. Nat Struct Mol Biol *11*, 394-403.

Hamdi, H. K., Nishio, H., Tavis, J., Zielinski, R., and Dugaiczyk, A. (2000). Alu-mediated phylogenetic novelties in gene regulation and development. J Mol Biol *299*, 931-939.

Hannon, G. J. (2002). RNA interference. Nature *418*, 244-251.

Hansen, J. J., Bross, P., Westergaard, M., Nielsen, M. N., Eiberg, H., Borglum, A. D., Mogensen, J., Kristiansen, K., Bolund, L., and Gregersen, N. (2003). Genomic structure of the human mitochondrial

chaperonin genes: HSP60 and HSP10 are localised head to head on chromosome 2 separated by a bidirectional promoter. Hum Genet *112*, 71-77.

Hardison, R., Xu, J., Jackson, J., Mansberger, J., Selifonova, O., Grotch, B., Biesecker, J., Petrykowska, H., and Miller, W. (1993). Comparative analysis of the locus control region of the rabbit beta-like gene cluster: HS3 increases transient expression of an embryonic epsilon-globin gene. Nucleic Acids Res *21*, 1265-1272.

Hardison, R. C., Roskin, K. M., Yang, S., Diekhans, M., Kent, W. J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., *et al.* (2003). Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res *13*, 13-26.

Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S., and Matsushima, K. (2004). 5'-end SAGE for the analysis of transcriptional start sites. Nat Biotechnol *22*, 1146-1149.

Helmlinger, D., Abou-Sleymane, G., Yvert, G., Rousseau, S., Weber, C., Trottier, Y., Mandel, J. L., and Devys, D. (2004). Disease progression despite early loss of polyglutamine protein expression in SCA7 mouse model. J Neurosci *24*, 1881-1887.

Helmlinger, D., Hardy, S., Abou-Sleymane, G., Eberlin, A., Bowman, A. B., Gansmuller, A., Picaud, S., Zoghbi, H. Y., Trottier, Y., Tora, L., and Devys, D. (2006). Glutamine-expanded ataxin-7 alters TFTC/STAGA recruitment and chromatin structure leading to photoreceptor dysfunction. PLoS Biol *4*, e67.

Hennig, S., Groth, D., and Lehrach, H. (2003). Automated Gene Ontology annotation for anonymous sequence data. Nucleic Acids Res *31*, 3712-3715.

Hermfisse, U., Schafer, D., Netzker, R., and Brand, K. (1996). The aldolase A promoter in proliferating rat thymocytes is regulated by a cluster of SP1 sites and a distal modulator. Biochem Biophys Res Commun *225*, 997-1005.

Herrera, R., Ro, H. S., Robinson, G. S., Xanthopoulos, K. G., and Spiegelman, B. M. (1989). A direct role for C/EBP and the AP-I-binding site in gene expression linked to adipocyte differentiation. Mol Cell Biol *9*, 5331-5339.

Higgins, D. G., and Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene *73*, 237-244.

Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., Bork, P., Burt, D. W., Groenen, M. A., Delany, M. E., *et al.* (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature *432*, 695-716.

Hogeveen, K. N., and Sassone-Corsi, P. (2006). Regulation of gene expression in post-meiotic male germ cells: CREM-signalling pathways and male fertility. Hum Fertil (Camb) *9*, 73-79.

Horwitz, K. B., Jackson, T. A., Bain, D. L., Richer, J. K., Takimoto, G. S., and Tung, L. (1996). Nuclear receptor coactivators and corepressors. Mol Endocrinol *10*, 1167-1177.

Howard, M. L., and Davidson, E. H. (2004). cis-Regulatory control circuits in development. Dev Biol *271*, 109-118.

Hsu, M. H., Savas, U., Griffin, K. J., and Johnson, E. F. (2001). Identification of peroxisome proliferator-responsive human genes by elevated expression of the peroxisome proliferator-activated receptor alpha in HepG2 cells. J Biol Chem *276*, 27950-27958.

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.

Iwama, H., and Gojobori, T. (2004). Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network. Proc Natl Acad Sci U S A *101*, 17156-17161.

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.

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.

Jeffery, C. J. (1999). Moonlighting proteins. Trends Biochem Sci *24*, 8-11.

Jeffery, C. J. (2003). Moonlighting proteins: old proteins learning new tricks. Trends Genet *19*, 415-417.

Jensen, L. J., Gupta, R., Staerfeldt, H. H., and Brunak, S. (2003). Prediction of human protein function according to Gene Ontology categories. Bioinformatics *19*, 635-642.

Jenuwein, T., and Allis, C. D. (2001). Translating the histone code. Science *293*, 1074-1080.

Jin, V. X., Singer, G. A., Agosto-Perez, F. J., Liyanarachchi, S., and Davuluri, R. V. (2006). Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. BMC Bioinformatics *7*, 114.

Jordan, I. K., Rogozin, I. B., Glazko, G. V., and Koonin, E. V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet *19*, 68-72.

Juge-Aubry, C., Pernin, A., Favez, T., Burger, A. G., Wahli, W., Meier, C. A., and Desvergne, B. (1997). DNA binding properties of peroxisome proliferator-activated receptor subtypes on various natural peroxisome proliferator response elements. Importance of the 5'-flanking region. J Biol Chem *272*, 25252-25259.

Kadonaga, J. T. (2002). The DPE, a core promoter element for transcription by RNA polymerase II. Exp Mol Med *34*, 259-264.

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

Kardassis, D., Falvey, E., Tsantili, P., Hadzopoulou-Cladaras, M., and Zannis, V. (2002). Direct physical interactions between HNF-4 and Sp1 mediate synergistic transactivation of the apolipoprotein CIII promoter. Biochemistry *41*, 1217-1228.

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.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J., *et al.* (2005). Antisense transcription in the mammalian transcriptome. Science *309*, 1564-1566.

Kawai, Y., Asai, K., Miura, Y., Inoue, Y., Yamamoto, M., Moriyama, A., Yamamoto, N., and Kato, T. (2003). Structure and promoter activity of the human glia maturation factor-gamma gene: a TATA-less, GC-rich and bidirectional promoter. Biochim Biophys Acta *1625*, 246-252.

Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res *31*, 3576-3579.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

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.

Keriel, A., Stary, A., Sarasin, A., Rochette-Egly, C., and Egly, J. M. (2002). XPD mutations prevent TFIIH-dependent transactivation by nuclear receptors and phosphorylation of RARalpha. Cell *109*, 125-135.

Khan, S., Situ, G., Decker, K., and Schmidt, C. J. (2003). GoFigure: automated Gene Ontology annotation. Bioinformatics *19*, 2484-2485.

Khatri, P., and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics *21*, 3587-3595.

Khorasanizadeh, S. (2004). The nucleosome: from genomic organization to genomic regulation. Cell *116*, 259-272.

Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H., *et al.* (2003). Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. Science *301*, 376-379.

Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. Nature *436*, 876-880.

Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S., and Hayashizaki, Y. (2003). Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. Genome Res *13*, 1324-1334.

Klingenhoff, A., Frech, K., Quandt, K., and Werner, T. (1999). Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. Bioinformatics *15*, 180-186.

Knudsen, S. (1999). Promoter2.0: for the recognition of PolII promoter sequences. Bioinformatics *15*, 356-361.

Kornberg, R. D., and Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell *98*, 285-294.

Koski, L. B., and Golding, G. B. (2001). The closest BLAST hit is often not the nearest neighbor. J Mol Evol *52*, 540-542.

Kotaja, N., Lin, H., Parvinen, M., and Sassone-Corsi, P. (2006). Interplay of PIWI/Argonaute protein MIWI and kinesin KIF17b in chromatoid bodies of male germ cells. J Cell Sci *119*, 2819-2825.

Kotaja, N., Macho, B., and Sassone-Corsi, P. (2005). Microtubule-independent and protein kinase A-mediated function of kinesin KIF17b controls the intracellular transport of activator of CREM in testis (ACT). J Biol Chem *280*, 31739-31745.

Kuersten, S., and Goodwin, E. B. (2003). The power of the 3' UTR: translational control and development. Nat Rev Genet *4*, 626-637.

Kumar, M., and Carmichael, G. G. (1998). Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. Microbiol Mol Biol Rev *62*, 1415-1434.

Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D., and Ebright, R. H. (1998). New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. Genes Dev *12*, 34-44.

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 *409*, 860-921.

Landry, J. R., Mager, D. L., and Wilhelm, B. T. (2003). Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet *19*, 640-648.

Lardenois, A., Chalmel, F., Bianchetti, L., Sahel, J. A., Leveillard, T., and Poch, O. (2006). PromAn: an integrated knowledge-based web server dedicated to promoter analysis. Nucleic Acids Res *34*, W578-583.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. Genomics *13*, 1095-1107.

Lavorgna, G., Sessa, L., Guffanti, A., Lassandro, L., and Casari, G. (2004). AntiHunter: searching BLAST output for EST antisense transcripts. Bioinformatics *20*, 583-585.

Leblanc, B. P., Benham, C. J., and Clark, D. J. (2000). An initiation element in the yeast CUP1 promoter is recognized by RNA polymerase II in the absence of TATA box-binding protein if the DNA is negatively supercoiled. Proc Natl Acad Sci U S A *97*, 10745-10750.

Lecompte, O., Thompson, J. D., Plewniak, F., Thierry, J., and Poch, O. (2001). Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene *270*, 17-30.

Lee, S. S., Pineau, T., Drago, J., Lee, E. J., Owens, J. W., Kroetz, D. L., Fernandez-Salguero, P. M., Westphal, H., and Gonzalez, F. J. (1995). Targeted disruption of the alpha isoform of the peroxisome proliferator-activated receptor gene in mice results in abolishment of the pleiotropic effects of peroxisome proliferators. Mol Cell Biol *15*, 3012-3022.

Lee, T. I., and Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. Annu Rev Genet *34*, 77-137.

Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N., and Wasserman, W. W. (2003). Identification of conserved regulatory elements by comparative genome analysis. J Biol *2*, 13.

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet *12*, 1725-1735.

Lettice, L. A., Horikoshi, T., Heaney, S. J., van Baren, M. J., van der Linde, H. C., Breedveld, G. J., Joosse, M., Akarsu, N., Oostra, B. A., Endo, N., *et al.* (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A *99*, 7548-7553.

Leveillard, T., Mohand-Said, S., Lorentz, O., Hicks, D., Fintz, A. C., Clerin, E., Simonutti, M., Forster, V., Cavusoglu, N., Chalmel, F., *et al.* (2004). Identification and characterization of rod-derived cone viability factor. Nat Genet *36*, 755-759.

Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature *424*, 147-151.

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian microRNA targets. Cell *115*, 787-798.

Lim, C. Y., Santoso, B., Boulay, T., Dong, E., Ohler, U., and Kadonaga, J. T. (2004). The MTE, a new core promoter element for transcription by RNA polymerase II. Genes Dev *18*, 1606-1617.

Lipman, D. J., and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. Science *227*, 1435-1441.

Liu, R., McEachin, R. C., and States, D. J. (2003). Computationally identifying novel NF-kappa B-regulated immune genes in the human genome. Genome Res *13*, 654-661.

Liu, Y., Liu, X. S., Wei, L., Altman, R. B., and Batzoglou, S. (2004). Eukaryotic regulatory element conservation analysis and identification using comparative genomics. Genome Res *14*, 451-458.

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.

Loots, G. G., and Ovcharenko, I. (2004). rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Res *32*, W217-221.

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.

Lopez, A. J. (1998). Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. Annu Rev Genet *32*, 279-305.

Lui, W. Y., Sze, K. L., and Lee, W. M. (2006). Nectin-2 expression in testicular cells is controlled via the functional cooperation between transcription factors of the Sp1, CREB, and AP-1 families. J Cell Physiol *207*, 144-157.

Macho, B., Brancorsini, S., Fimia, G. M., Setou, M., Hirokawa, N., and Sassone-Corsi, P. (2002). CREM-dependent transcription in male germ cells controlled by a kinesin. Science *298*, 2388-2390.

Maclean, J. A., 2nd, and Wilkinson, M. F. (2005). Gene regulation in spermatogenesis. Curr Top Dev Biol *71*, 131-197.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res *33*, D54-58.

Mangelsdorf, D. J., Thummel, C., Beato, M., Herrlich, P., Schutz, G., Umesono, K., Blumberg, B., Kastner, P., Mark, M., Chambon, P., and Evans, R. M. (1995). The nuclear receptor superfamily: the second decade. Cell *83*, 835-839.

Margulies, E. H., Vinson, J. P., Miller, W., Jaffe, D. B., Lindblad-Toh, K., Chang, J. L., Green, E. D., Lander, E. S., Mullikin, J. C., and Clamp, M. (2005). An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci U S A *102*, 4795-4800.

Marino-Ramirez, L., Spouge, J. L., Kanga, G. C., and Landsman, D. (2004). Statistical analysis of over-represented words in human promoter sequences. Nucleic Acids Res *32*, 949-958.

Mathew, S., Mascareno, E., and Siddiqui, M. A. (2004). A ternary complex of transcription factors, Nished and NFATc4, and co-activator p300 bound to an intronic sequence, intronic regulatory element, is pivotal for the up-regulation of myosin light chain-2v gene in cardiac hypertrophy. J Biol Chem *279*, 41018-41027.

Matis, S., Xu, Y., Shah, M., Guan, X., Einstein, J. R., Mural, R., and Uberbacher, E. (1996). Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence. Comput Chem *20*, 135-140.

Matthaei, J. H., Jones, O. W., Martin, R. G., and Nirenberg, M. W. (1962). Characteristics and composition of RNA coding units. Proc Natl Acad Sci U S A *48*, 666-677.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., *et al.* (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res *34*, D108-110.

Maxam, A. M., and Gilbert, W. (1977). A new method for sequencing DNA. Proc Natl Acad Sci U S A *74*, 560-564.

Maxson, R., Cohn, R., Kedes, L., and Mohun, T. (1983). Expression and organization of histone genes. Annu Rev Genet *17*, 239-277.

McClelland, M., Florea, L., Sanderson, K., Clifton, S. W., Parkhill, J., Churcher, C., Dougan, G., Wilson, R. K., and Miller, W. (2000). Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi. Nucleic Acids Res *28*, 4974-4986.

McEwen, G. K., Woolfe, A., Goode, D., Vavouri, T., Callaway, H., and Elgar, G. (2006). Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Res *16*, 451-465.

McKusick, V. A., and Ruddle, F. H. (1987). Toward a complete map of the human genome. Genomics *1*, 103-106.

Mears, A. J., Kondo, M., Swain, P. K., Takada, Y., Bush, R. A., Saunders, T. L., Sieving, P. A., and Swaroop, A. (2001). Nrl is required for rod photoreceptor development. Nat Genet *29*, 447-452.

Meyer, S., Temme, C., and Wahle, E. (2004). Messenger RNA turnover in eukaryotes: pathways and enzymes. Crit Rev Biochem Mol Biol *39*, 197-216.

Mighell, A. J., Smith, N. R., Robinson, P. A., and Markham, A. F. (2000). Vertebrate pseudogenes. FEBS Lett *468*, 109-114.

Mitchell, P. J., and Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science *245*, 371-378.

Mitton, K. P., Swain, P. K., Chen, S., Xu, S., Zack, D. J., and Swaroop, A. (2000). The leucine zipper of NRL interacts with the CRX homeodomain. A possible mechanism of transcriptional synergy in rhodopsin regulation. J Biol Chem *275*, 29794-29799.

Mohand-Said, S., Deudon-Combe, A., Hicks, D., Simonutti, M., Forster, V., Fintz, A. C., Leveillard, T., Dreyfus, H., and Sahel, J. A. (1998). Normal retina releases a diffusible factor stimulating cone survival in the retinal degeneration mouse. Proc Natl Acad Sci U S A *95*, 8357-8362.

Momota, R., Sugimoto, M., Oohashi, T., Kigasawa, K., Yoshioka, H., and Ninomiya, Y. (1998). Two genes, COL4A3 and COL4A4 coding for the human alpha3(IV) and alpha4(IV) collagen chains are arranged head-to-head on chromosome 2q36. FEBS Lett *424*, 11-16.

Morgenstern, B., Frech, K., Dress, A., and Werner, T. (1998). DIALIGN: finding local similarities by multiple sequence alignment. Bioinformatics *14*, 290-294.

Munroe, S. H. (2004). Diversity of antisense regulation in eukaryotes: multiple mechanisms, emerging patterns. J Cell Biochem *93*, 664-671.

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.

Ng, P., Wei, C. L., Sung, W. K., Chiu, K. P., Lipovich, L., Ang, C. C., Gupta, S., Shahab, A., Ridwan, A., Wong, C. H., *et al.* (2005). Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods *2*, 105-111.

Nobrega, M. A., Ovcharenko, I., Afzal, V., and Rubin, E. M. (2003). Scanning human gene deserts for long-range enhancers. Science *302*, 413.

Noguchi, M., Miyamoto, S., Silverman, T. A., and Safer, B. (1994). Characterization of an antisense Inr element in the eIF-2 alpha gene. J Biol Chem *269*, 29161-29167.

Norris, J., Fan, D., Aleman, C., Marks, J. R., Futreal, P. A., Wiseman, R. W., Iglehart, J. D., Deininger, P. L., and McDonnell, D. P. (1995). Identification of a new subclass of Alu DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. J Biol Chem *270*, 22777-22782.

Novina, C. D., and Roy, A. L. (1996). Core promoters and transcriptional control. Trends Genet *12*, 351-355.

Ohler, U., Liao, G. C., Niemann, H., and Rubin, G. M. (2002). Computational analysis of core promoters in the Drosophila genome. Genome Biol *3*, RESEARCH0087.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., *et al.* (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563-573.

Ovcharenko, I., Boffelli, D., and Loots, G. G. (2004). eShadow: a tool for comparing closely related sequences. Genome Res *14*, 1191-1198.

Ovcharenko, I., and Nobrega, M. A. (2005). Identifying synonymous regulatory elements in vertebrate genomes. Nucleic Acids Res *33*, W403-407.

Pesole, G., Bernardi, G., and Saccone, C. (1999). Isochore specificity of AUG initiator context of human genes. FEBS Lett *464*, 60-62.

Peterson, M. G. (1988). DNA sequencing using Taq polymerase. Nucleic Acids Res *16*, 10915.

Phelan, J. K., and Bok, D. (2000). A brief review of retinitis pigmentosa and the identified retinitis pigmentosa genes. Mol Vis *6*, 116-124.

Pittler, S. J., Zhang, Y., Chen, S., Mears, A. J., Zack, D. J., Ren, Z., Swain, P. K., Yao, S., Swaroop, A., and White, J. B. (2004). Functional analysis of the rod photoreceptor cGMP phosphodiesterase alpha-subunit gene promoter: Nrl and Crx are required for full transcriptional activity. J Biol Chem *279*, 19800-19807.

Plewniak, F., Bianchetti, L., Brelivet, Y., Carles, A., Chalmel, F., Lecompte, O., Mochel, T., Moulinier, L., Muller, A., Muller, J., *et al.* (2003). PipeAlign: A new toolkit for protein family analysis. Nucleic Acids Res *31*, 3829-3832.

Plewniak, F., Thompson, J. D., and Poch, O. (2000). Ballast: blast post-processing based on locally conserved segments. Bioinformatics *16*, 750-759.

Pohar, T. T., Sun, H., and Davuluri, R. V. (2004). HemoPDB: Hematopoiesis Promoter Database, an information resource of transcriptional regulation in blood cell development. Nucleic Acids Res *32*, D86-90.

Ponger, L., Duret, L., and Mouchiroud, D. (2001). Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res *11*, 1854-1860.

Ponger, L., and Mouchiroud, D. (2002). CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics *18*, 631-633.

Portera-Cailliau, C., Sung, C. H., Nathans, J., and Adler, R. (1994). Apoptotic photoreceptor cell death in mouse models of retinitis pigmentosa. Proc Natl Acad Sci U S A *91*, 974-978.

Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E. M., Couronne, O., and Pennacchio, L. A. (2006). Close sequence comparisons are sufficient to identify human cis-regulatory elements. Genome Res *16*, 855-863.

Prestridge, D. S. (1996). SIGNAL SCAN 4.0: additional databases and sequence formats. Comput Appl Biosci *12*, 157-160.

Prince, V. E., and Pickett, F. B. (2002). Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet *3*, 827-837.

Qian, J., Esumi, N., Chen, Y., Wang, Q., Chowers, I., and Zack, D. J. (2005). Identification of regulatory targets of tissue-specific transcription factors: application to retina-specific gene regulation. Nucleic Acids Res *33*, 3479-3491.

Qiu, P. (2003). Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. Biochem Biophys Res Commun *309*, 495-501.

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.

Qvist, H., Sjostrom, H., and Noren, O. (1998). The TATA-less, GC-rich porcine dipeptidylpeptidase IV (DPPIV) promoter shows bidirectional activity. Biol Chem *379*, 75-81.

Reinberg, D., Orphanides, G., Ebright, R., Akoulitchev, S., Carcamo, J., Cho, H., Cortes, P., Drapkin, R., Flores, O., Ha, I., *et al.* (1998). The RNA polymerase II general transcription factors: past, present, and future. Cold Spring Harb Symp Quant Biol *63*, 83-103.

Renaud, J. P., and Moras, D. (2000). Structural studies on nuclear receptors. Cell Mol Life Sci *57*, 1748-1769.

Richmond, T. J., and Davey, C. A. (2003). The structure of DNA in the nucleosome core. Nature *423*, 145-150.

Rival, Y., Stennevin, A., Puech, L., Rouquette, A., Cathala, C., Lestienne, F., Dupont-Passelaigue, E., Patoiseau, J. F., Wurch, T., and Junquero, D. (2004). Human adipocyte fatty acid-binding protein (aP2) gene promoter-driven reporter assay discriminates nonlipogenic peroxisome proliferator-activated receptor gamma ligands. J Pharmacol Exp Ther *311*, 467-475.

Robinson-Rechavi, M., Escriva Garcia, H., and Laudet, V. (2003). The nuclear receptor superfamily. J Cell Sci *116*, 585-586.

Rockman, M. V., and Wray, G. A. (2002). Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol *19*, 1991-2004.

Rodriguez, J. C., Gil-Gomez, G., Hegardt, F. G., and Haro, D. (1994). Peroxisome proliferator-activated receptor mediates induction of the mitochondrial 3-hydroxy-3-methylglutaryl-CoA synthase gene by fatty acids. J Biol Chem *269*, 18767-18772.

Roeder, R. G. (1996). The role of general initiation factors in transcription by RNA polymerase II. Trends Biochem Sci *21*, 327-335.

Rosenberg, M. S. (2005). Evolutionary distance estimation and fidelity of pair wise sequence alignment. BMC Bioinformatics *6*, 102.

Ross, C. A. (2002). Polyglutamine pathogenesis: emergence of unifying mechanisms for Huntington's disease and related disorders. Neuron *35*, 819-822.

Sadiq, M., Hildebrandt, M., Maniak, M., and Nellen, W. (1994). Developmental regulation of antisense-mediated gene silencing in Dictyostelium. Antisense Res Dev *4*, 263-267.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., and Lenhard, B. (2004a). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res *32*, D91-94.

Sandelin, A., Bailey, P., Bruce, S., Engstrom, P. G., Klos, J. M., Wasserman, W. W., Ericson, J., and Lenhard, B. (2004b). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics *5*, 99.

Sandelin, A., and Wasserman, W. W. (2004). Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. J Mol Biol *338*, 207-215.

Sanger, F., Coulson, A. R., Friedmann, T., Air, G. M., Barrell, B. G., Brown, N. L., Fiddes, J. C., Hutchison, C. A., 3rd, Slocombe, P. M., and Smith, M. (1978). The nucleotide sequence of bacteriophage phiX174. J Mol Biol *125*, 225-246.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A *74*, 5463-5467.

Sanger, F., and Thompson, E. O. (1953). The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem J *53*, 353-366.

Satchwell, S. C., Drew, H. R., and Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. J Mol Biol *191*, 659-675.

Sauer, T., Shelest, E., and Wingender, E. (2006). Evaluating phylogenetic footprinting for human-rodent comparisons. Bioinformatics *22*, 430-437.

Scherf, M., Klingenhoff, A., and Werner, T. (2000). Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol *297*, 599-606.

Schmid, C. D., Perier, R., Praz, V., and Bucher, P. (2006). EPD in its twentieth year: towards complete promoter coverage of selected model organisms. Nucleic Acids Res *34*, D82-85.

Schmidt, C., Fischer, G., Kadner, H., Genersch, E., Kuhn, K., and Poschl, E. (1993). Differential effects of DNA-binding proteins on bidirectional transcription from the common promoter region of human collagen type IV genes COL4A1 and COL4A2. Biochim Biophys Acta *1174*, 1-10.

Schneider, T. D., and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res *18*, 6097-6100.

Schug, J., Schuller, W. P., Kappen, C., Salbaum, J. M., Bucan, M., and Stoeckert, C. J., Jr. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol *6*, R33.

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.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., Wang, J. P., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772-778.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., *et al.* (2002). Functional annotation of a full-length Arabidopsis cDNA collection. Science *296*, 141-145.

Shapiro, J. A. (2005). A 21st century view of evolution: genome system architecture, repetitive DNA, and natural genetic engineering. Gene *345*, 91-100.

Siddiqui, A. S., Khattra, J., Delaney, A. D., Zhao, Y., Astell, C., Asano, J., Babakaiff, R., Barber, S., Beland, J., Bohacec, S., *et al.* (2005). A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. Proc Natl Acad Sci U S A *102*, 18485-18490.

Siep, M., Sleddens-Linkels, E., Mulders, S., van Eenennaam, H., Wassenaar, E., Van Cappellen, W. A., Hoogerbrugge, J., Grootegoed, J. A., and Baarends, W. M. (2004). Basic helix-loop-helix transcription factor Tcfl5 interacts with the Calmegin gene promoter in mouse spermatogenesis. Nucleic Acids Res *32*, 6425-6436.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034-1050.

Smale, S. T., and Kadonaga, J. T. (2003). The RNA polymerase II core promoter. Annu Rev Biochem *72*, 449-479.

Smit, A. F. (1996). The origin of interspersed repeats in the human genome. Curr Opin Genet Dev *6*, 743-748.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. J Mol Biol *147*, 195-197.

Solovyev, V. V., and Shahmuradov, I. A. (2003). PromH: Promoters identification using orthologous genomic sequences. Nucleic Acids Res *31*, 3540-3545.

Stalker, J., Gibbins, B., Meidl, P., Smith, J., Spooner, W., Hotz, H. R., and Cox, A. V. (2004). The Ensembl Web site: mechanics of a genome browser. Genome Res *14*, 951-955.

Stapleton, M., Liao, G., Brokstein, P., Hong, L., Carninci, P., Shiraki, T., Hayashizaki, Y., Champe, M., Pacleb, J., Wan, K., *et al.* (2002). The Drosophila gene collection: identification of putative full-length cDNAs for 70% of D. melanogaster genes. Genome Res *12*, 1294-1300.

Stoeckert, C. J., Jr., Salas, F., Brunk, B., and Overton, G. C. (1999). EpoDB: a prototype database for the analysis of genes expressed during vertebrate erythropoiesis. Nucleic Acids Res *27*, 200-203.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*, 16-23.

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., *et al.* (2000). Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen. Nature *406*, 959-964.

Su, W., Jackson, S., Tjian, R., and Echols, H. (1991). DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1. Genes Dev *5*, 820-826.

Sugimoto, M., Oohashi, T., and Ninomiya, Y. (1994). The genes COL4A5 and COL4A6, coding for basement membrane collagen chains alpha 5(IV) and alpha 6(IV), are located head-to-head in close proximity on human chromosome Xq22 and COL4A6 is transcribed from two alternative promoters. Proc Natl Acad Sci U S A *91*, 11679-11683.

Suter, B., Schnappauf, G., and Thoma, F. (2000). Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. Nucleic Acids Res *28*, 4083-4089.

Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y., *et al.* (2001). Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res *11*, 677-684.

Suzuki, Y., Yamashita, R., Nakai, K., and Sugano, S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res *30*, 328-331.

Suzuki, Y., Yamashita, R., Shirota, M., Sakakibara, Y., Chiba, J., Mizushima-Sugano, J., Nakai, K., and Sugano, S. (2004a). Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. Genome Res *14*, 1711-1718.

Suzuki, Y., Yamashita, R., Sugano, S., and Nakai, K. (2004b). DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. Nucleic Acids Res *32*, D78-81.

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.

Taylor, M. S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Semple, C. A. (2006). Heterotachy in mammalian promoter evolution. PLoS Genet *2*, e30.

Thomas, M. C., and Chiang, C. M. (2006). The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol *41*, 105-178.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res *25*, 4876-4882.

Thompson, J. D., Higgins, D. G., 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.

Thompson, J. D., Plewniak, F., Ripp, R., Thierry, J. C., and Poch, O. (2001). Towards a reliable objective function for multiple sequence alignments. J Mol Biol *314*, 937-951.

Thompson, J. D., Plewniak, F., Thierry, J., and Poch, O. (2000). DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Res *28*, 2919-2926.

Thompson, J. D., Prigent, V., and Poch, O. (2004a). LEON: multiple aLignment Evaluation Of Neighbours. Nucleic Acids Res *32*, 1298-1307.

Thompson, J. D., Thierry, J. C., and Poch, O. (2003). RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics *19*, 1155-1161.

Thompson, W., Palumbo, M. J., Wasserman, W. W., Liu, J. S., and Lawrence, C. E. (2004b). Decoding human regulatory circuits. Genome Res *14*, 1967-1974.

Tjian, R., and Maniatis, T. (1994). Transcriptional activation: a complex puzzle with few easy pieces. Cell *77*, 5-8.

Tontonoz, P., Hu, E., and Spiegelman, B. M. (1995). Regulation of adipocyte gene expression and differentiation by peroxisome proliferator activated receptor gamma. Curr Opin Genet Dev *5*, 571-576.

Torrents, D., Suyama, M., Zdobnov, E., and Bork, P. (2003). A genome-wide survey of human pseudogenes. Genome Res *13*, 2559-2567.

Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P., and Myers, R. M. (2004). An abundance of bidirectional promoters in the human genome. Genome Res *14*, 62-66.

Trinklein, N. D., Aldred, S. J., Saldanha, A. J., and Myers, R. M. (2003). Identification and functional analysis of human transcriptional promoters. Genome Res *13*, 308-312.

Ureta-Vidal, A., Ettwiller, L., and Birney, E. (2003). Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet *4*, 251-262.

Valdar, W. S. (2002). Scoring residue conservation. Proteins *48*, 227-241.

Vanin, E. F. (1985). Processed pseudogenes: characteristics and evolution. Annu Rev Genet *19*, 253-272.

Vansant, G., and Reynolds, W. F. (1995). The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. Proc Natl Acad Sci U S A *92*, 8229-8233.

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.

Versteeg, R., van Schaik, B. D., van Batenburg, M. F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H. J., and van Kampen, A. H. (2003). The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res *13*, 1998-2004.

Vinogradov, A. E. (2003). Isochores and tissue-specificity. Nucleic Acids Res *31*, 5212-5220.

Vinogradov, A. E. (2004). Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet *20*, 248-253.

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*, D95-97.

Wahli, W., Braissant, O., and Desvergne, B. (1995). Peroxisome proliferator activated receptors: transcriptional regulators of adipogenesis, lipid metabolism and more. Chem Biol *2*, 261-266.

Wall, D. P., Fraser, H. B., and Hirsh, A. E. (2003). Detecting putative orthologs. Bioinformatics *19*, 1710-1711.

Walsh, S., and Barrell, B. (1996). The Saccharomyces cerevisiae genome on the World Wide Web. Trends Genet *12*, 276-277.

Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., Samudrala, R., Yu, J., and Wong, G. K. (2004). Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. Nature *431*, 1 p following 757; discussion following 757.

Wang, Q., Carroll, J. S., and Brown, M. (2005). Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. Mol Cell *19*, 631-642.

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.

Wasserman, W. W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet *5*, 276-287.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Watson, J. D., and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Werner, A., and Berdal, A. (2005). Natural antisense transcripts: sound or silence? Physiol Genomics *23*, 125-131.

Werner, T. (1999). Models for prediction and recognition of eukaryotic promoters. Mamm Genome *10*, 168-175.

Werner, T. (2002). Finding and decrypting of promoters contributes to the elucidation of gene function. In Silico Biol *2*, 249-255.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., *et al.* (2006). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res *34*, D173-180.

Wicker, N., Dembele, D., Raffelsberger, W., and Poch, O. (2002). Density of points clustering, application to transcriptomic data analysis. Nucleic Acids Res *30*, 3992-4000.

Widom, J. (2001). Role of DNA sequence in nucleosome stability and dynamics. Q Rev Biophys *34*, 269-324.

Wilusz, C. J., and Wilusz, J. (2004). Bringing the role of mRNA decay in the control of gene expression into focus. Trends Genet *20*, 491-497.

Wingender, E. (1988). Compilation of transcription regulating proteins. Nucleic Acids Res *16*, 1879-1902.

Wingender, E. (1997). [Classification of eukaryotic transcription factors]. Mol Biol (Mosk) *31*, 584-600.

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.

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.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A *87*, 4576-4579.

Wolffe, A. P. (1994). Transcriptional activation. Switched-on chromatin. Curr Biol *4*, 525-528.

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.

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.

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., *et al.* (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res *34*, D187-191.

Wu, Y., Zhang, Y., and Zhang, J. (2005). Distribution of exonic splicing enhancer elements in human genes. Genomics *86*, 329-336.

Wyrick, J. J., Holstege, F. C., Jennings, E. G., Causton, H. C., Shore, D., Grunstein, M., Lander, E. S., and Young, R. A. (1999). Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. Nature *402*, 418-421.

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.

Xuan, Z., Zhao, F., Wang, J., Chen, G., and Zhang, M. Q. (2005). Genome-wide promoter extraction and analysis in human, mouse, and rat. Genome Biol *6*, R72.

Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K. (2005). Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. Gene *350*, 129-136.

Yamashita, R., Suzuki, Y., Wakaguri, H., Tsuritani, K., Nakai, K., and Sugano, S. (2006). DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. Nucleic Acids Res *34*, D86-89.

Ye, L., and Huang, X. (2005). MAP2: multiple alignment of syntenic genomic sequences. Nucleic Acids Res *33*, 162-170.

Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., *et al.* (2003). Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol *21*, 379-386.

Yuh, C. H., Dorman, E. R., Howard, M. L., and Davidson, E. H. (2004). An otx cis-regulatory module: a key node in the sea urchin endomesoderm gene regulatory network. Dev Biol *269*, 536-551.

Zehetner, G. (2003). OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. Nucleic Acids Res *31*, 3799-3803.

Zerucha, T., Stuhmer, T., Hatch, G., Park, B. K., Long, Q., Yu, G., Gambarotta, A., Schultz, J. R., Rubenstein, J. L., and Ekker, M. (2000). A highly conserved enhancer in the Dlx5/Dlx6 intergenic

region is the site of cross-regulatory interactions between Dlx genes in the embryonic forebrain. J Neurosci *20*, 709-721.

Zhang, Z., Carriero, N., and Gerstein, M. (2004). Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet *20*, 62-67.

Zhang, Z., and Gerstein, M. (2003). Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J Biol *2*, 11.

Zoghbi, H. Y., and Orr, H. T. (2000). Glutamine repeats and neurodegeneration. Annu Rev Neurosci *23*, 217-247.

Annexes

Annexe 1 - TRANSFAC PWM and matrix entry.

Figure 103 shows an example of such a matrix, for the human transcription factor GATA-1, from the widely used TRANSFAC database (Wingender et al., 2001).

TRANSFAC accession number: M00127
TRANSFAC identifier: VSGATA1_03
Name: GATA-1
Description: GATA-binding factor 1

Position	A	C	G	T	Consensus sequence
1	4	1	2	0	R
2	1	1	3	2	N
3	1	2	4	0	S
4	2	2	2	1	N
5	3	0	2	2	D
6	0	0	12	0	G
7	12	0	0	0	A
8	0	0	0	12	T
9	12	0	0	0	A
10	8	1	3	0	A
11	1	4	4	3	N
12	3	4	3	2	N
13	3	1	7	1	G
14	2	4	4	2	N

Statistical basis: 12 selected binding sequences

Figure 103. Position Weight Matrix (PWM) for the human transcription factor GATA-1. Extracted from Zhang *et al.*(Zhang and Gerstein, 2003).

The format of the Transfac matrix entries is illustrated in Figure 104.

```
AC        Accession no.
XX
ID        Identifier
XX
DT        Date; author
XX
NA        Name of the binding factor
XX
DE        Short factor description
XX
BF        List of linked factor entries
XX

PO        A   C   G   T      Position within the aligned sequences,
01                           frequency of A, C, G, T residues, resp.;
02                           last column: deduced consensus in
03                           IUPAC 15-letter code
XX
BA        Statistical basis
XX
BS        Factor binding sites underlying the matrix
BS        (SITE accession no.; Start position for matrix sequence; length of sequence used;
BS        number of gaps inserted; strand orientation)
XX
CC        Comments
XX
RX        MEDLINE ID
RN        Reference no.
RA        Reference authors
RT        Reference title
RL        Reference data
XX
//
```

Figure 104. Format of a TRANSFAC matrix entry.

The matrix entries have an identifier that indicates one of six groups of biological species (V$, vertebrates; I$, insects; P$, plants; F$, fungi; N$, nematodes; B$, bacteria), followed by an acronym for the factor the matrix refers to, and a consecutive number discriminating between different matrices for the same factor. Thus as an example, V$OCT1_02 indicates the second matrix for vertebral Oct-1 factor. Instead of the consecutive number, those matrices which have been generated from TRANSFAC® SITE entries connected to a certain transcription factor, IDs end with an abbreviation of the least quality of the sites used to construct the matrix. E. g., V$CREB_Q2 is a matrix constructed of CREB binding sites of quality 2 or better. Finally, a matrix with an ID like V$AP1_C has been derived from a "consensus description". The matrix area gives the nucleotide frequencies observed in aligned binding sites of the corresponding transcription factor (or, more generally, in aligned sites of the described function). An example of a Transfac matrix is shown in Figure 103. An additional column at the right of the matrix area depicts the IUPAC string consensus derived from the matrix according to the following rules (adapted from (Cavener, 1987)): a single nucleotide is shown if its frequency is greater than 50% and at least twice as high as the second most frequent nucleotide.

Annexe 2 - Brief description of programs used for noncoding DNA alignment.

ClustalW (Thompson et al., 1994) is a progressive multiple alignment tool. ClustalW incorporates the global dynamic programming algorithm developed by Needleman and Wunsch, 1971 (see section 1.6.2.1). The approach of progressive alignment constructs a multiple alignment by successive applications of a pairwise alignment algorithm. ClustalW is widely used for protein sequence alignments but in some cases, it is also applied for promoter sequence alignment. In this context, ClustalW gives good results if the regions to align are highly conserved. Otherwise, it does not lead to promoter optimized alignments (Sauer et al., 2006). Indeed, Rosenberg found that if less than 50% of positions between two non-coding sequences are identical, ClustalW creates pairwise alignments that do not differ from random data (Rosenberg, 2005). This can be explained by the fact that Clustal programs have been designed primarily for aligning multiple protein sequences.

DIALIGN (Morgenstern et al., 1998) is a segment-to-segment alignment algorithm that chains locally conserved blocks between several sequences into a multiple alignment. Like the BLAST algorithms, DIALIGN looks for short ungapped segments having a similarity that deviates from what would be expected by random chance, keeping segments with a score above a certain threshold. These high scoring segments are then aligned into a collinear global alignment using a dynamic programming algorithm.

The CHAOS program (Brudno et al., 2003a) is a pairwise local alignment tool. CHAOS (Chains Of Seeds) works by chaining pairs of seeds, similar regions in each of the two input sequences. Its algorithm detects local alignments using multiple short inexact words. Chaos starts by finding all words between the two sequences of a specified length and a specified maximum number of mismatches. These words are then chained together if they are close together in both sequences. These maximal chains are then scored and all chains that are above a specified threshold are reported.

LAGAN (Limited Area Global Alignment of Nucleotides) (Brudno et al., 2003b) is a pairwise aligner. LAGAN aligns two genomic sequences in three main steps: (1) generation of local alignments between the two sequences, (2) construction of a rough global map by chaining an ordered subset of the local alignments, and (3) computation of the final global alignment, by finding the best alignment that remains within a limited area around the rough global map. For the first step, LAGAN uses the CHAOS algorithm (Brudno and Morgenstern 2002).

MLAGAN (Multi-LAGAN) (Brudno et al., 2003b) is a multiple aligner based on progressive alignment with LAGAN. It aligns genomic sequences in two main phases: (1) a progressive alignment phase that constructs a multiple alignment by successively aligning two

sequences, or intermediate multiple alignment, with the LAGAN algorithm, and (2) an optional iterative improvement phase that successively removes each sequence from the multiple alignment, and realigns it to the rest of the alignment, until no significant improvement is observed.

AVID (Bray et al., 2003) is a pairwise global alignment program whose general strategy for aligning two sequences is to anchor and align iteratively. A set of maximal (but not necessarily unique) exact matches between the sequences is constructed using a suffix tree. Dynamic programming is then used to order and orient the longest matches, which are then fixed. For each remaining subsequence between the fixed matches, the process is repeated until every base is aligned. The multiple alignment program MAVID (Gibbs et al., 2004) involves the progressive alignment of ancestor sequences (inferred using maximum-likelihood estimation) along a phylogenetic guide tree.

BLASTZ (Schwartz et al., 2003) is a pairwise local alignment tool that is based on the Gapped BLAST algorithm (Altschul et al., 1997) that has been redesigned for the alignment of long genomic sequences. BLASTZ first removes lineage-specific interspersed repeats from each sequence, then searches for short near-perfect matches between the two sequences. Each match is extended first using gap-free dynamic programming and if it scores above a specified threshold it will be extended using dynamic programming with gaps; extended matches that score above a specified threshold are then kept. Part of the unique implementation of BLASTZ is that it can be forced to return alignments that are both unique within each sequence as well as collinear with respect to each other.

The general problem addressed by MULTIZ (Blanchette et al., 2004) is to align two blocksets generated with the BLASTZ program and with different reference sequences. MULTIZ aligns the segments of sequences to each other by dynamic programming. This program creates "reference sequence" alignements, that is, a sequence is fixed as the reference to which all other sequences are compared. MULTIZ is the component of the TBA program (see below) that does the dynamic-programming alignment step. MULTIZ can be used with sequences that are fragmented or have rearrangements such as inversions and duplications. This program is used to build whole-genome alignments for the UCSC Genome Browser (Kent et al., 2002).

TBA (Threaded Blocksets Aligner) (Blanchette et al., 2004) is a local alignment program. TBA is implemented as a suite of independently executing programs. It produces set of blocks under the assumption that the matching regions occur in the same order and orientation in all species. The "blocks" are defined as local alignments of some or all of the given sequences. Each position in the given sequences appears precisely once in the blocks created

by TBA. The multiple sequences are aligned in the order of their pairwise distance in the phylogenetic tree. This program does not require a "reference sequence".

Annexe 3 - Brief description of the main phylogenetic footprinting programs.

rVISTA (regulatory VISTA) (Loots et al., 2002) identifies TFBS matches in the individual sequences, then identifies theglobally aligned noncoding TFBSs. Finally, it calculates local conservation extended upstream and downstream of the TFBSs. Local conservation of at least 80% sequence identity in a 20 bp sliding window spanning the binding site indicates aligned-and-conserved TFBSs. Finally a graphical display is created to visualize the conservation profile of the genomic locus and the TFBSs.

rVISTA 2.0 (Loots and Ovcharenko, 2004) uses the same strategy as rVISTA but can process alignments generated by both the zPicture and blastz alignment programs or use pre-computed pairwise alignments of several vertebrate genomes available from the ECR Browser and GALA database.

CONSITE (Lenhard et al., 2003) aligns input promoter sequences, calculates the degree of conservation in the alignment and scans both promoters for TFBSs. Finally CONSITE reports the putative TFBSs that are both situated in conserved regions and located as pairs of sites in equivalent positions in alignments between two orthologous sequences.

CONREAL (conserved regulatory elements anchored alignments) (Berezikov et al., 2004) does not depend on prior alignment of orthologous promoter sequences. All TFBSs are first predicted on each orthologous promoter. Then, binding sites for the same TF are anchored between the orthologous promoters, starting with the TFBSs with the highest scores and assuming collinear conservation of the binding sites. This program performs as well as other approaches depending on prior alignment when applied to the comparison of sequences from human and rodent. This algorithm is more useful for aligning promoter elements of more divergent species such as human and Fugu.

FOOTER (Corcoran et al., 2005) combines two statistics in order to score a pair of putative regulatory sites. Both promoter sequences are scanned and foreach TF the top scoring sites are retained in each promoter. By default, FOOTER retains one top scoring motif per TF per 300 bp of promoter sequence. Then, all sites are compared in order to find the best matching pairs according to the position conservation of the two motifs in the sequence and according to the agreement of the motifs with the corresponding PSSM model. As a final result, FOOTER provides a list of predicted TFBSs in a table format.

Annexe 4 - Publication 1.

Annexe 5 - Publication 2.

Annexe 6 - Publication 3.

Polyglutamine expansion causes neurodegeneration by altering the neuronal differentiation program

Gretta Abou-Sleymane, Frédéric Chalmel, Dominique Helmlinger, **Aurélie Lardenois,** Christelle Thibault, Chantal Weber, Karine Mérienne, Jean-Louis Mandel, Olivier Poch, Didier Devys et Yvon Trottier

Human Molecular Genetics, 2006, Vol. 15, Pages 691-703

Pages 691 à 703 :

Abstract

The exponential accumulation of high-throughput experimental data and complete genome sequences has greatly encouraged promoter sequence analysis through bioinformatics. To date, bioinformatics approaches have been used by biologists to facilitate the identification of regulatory motifs in promoter regions before engaging in time-consuming biochemical characterizations. However, the emergence of a huge amount of experimental data, prediction programs and complementary methods means that integrative approaches have become essential to improve *in silico* promoter analysis.

In this context, we have developed PromAn, a versatile and integrative tool which provides a wide range of state-of-the-art promoter analyses. The program requires minimal prior knowledge of the input genomic sequence and includes an evaluation of the evolutionary conservation of promoter regions, a validation of the transcriptional start sites as well as a prediction of potentially active transcription factor binding sites. PromAn has been implemented in two expert-guided versions (local and web server) as well as a high-throughput automatic version that is used in combination with gene groups assumed to be co-regulated.

In the context of a number of collaborations with different research groups, the efficiency of PromAn has been demonstrated in strong synergy with experimental validations through the localization and identification of *bona-fide* transcription factor binding sites. Hopefully, the PromAn high-throughput version will facilitate the understanding of complete regulatory networks and their impact in human health and diseases.

Résumé

L'accumulation exponentielle de données expérimentales générées par les technologies à haut débit a considérablement favorisé l'analyse bioinformatique des séquences promotrices. A ce jour, des approches bioinformatiques ont été utilisées par les biologistes afin de faciliter l'identification des motifs de régulation dans les régions promotrices avant d'entreprendre des caractérisations biochimiques onéreuses en temps. Cependant, l'émergence d'une quantité colossale de données expérimentales, de programmes de prédiction et de méthodes complémentaires souligne l'absolue nécessité de développer des approches intégratives afin d'améliorer l'analyse des promoteurs assistée par ordinateur.

Dans ce contexte, nous avons développé PromAn, un outil polyvalent et intégratif qui offre un panel de modules couvrant une grande partie des approches utilisées dans le domaine de l'analyse des promoteurs. Le programme ne requiert aucune connaissance préalable sur la séquence génomique à étudier et inclut une évaluation de la conservation au cours de l'évolution des régions promotrices, une validation des sites d'initiation de la transcription ainsi qu'une prédiction des sites de fixation de facteurs de transcription potentiellement actifs. PromAn implémente deux versions semi-automatiques (en local et sur un serveur web) ainsi qu'une version automatisée dédiée aux analyses à haut débit et utilisée en étroite conjonction avec des groupes de gènes potentiellement co-régulés.

Dans le cadre de nombreuses collaborations avec divers groupes de recherche, l'efficacité de PromAn a pu être démontrée en étroite synergie avec des validations expérimentales afin de localiser et d'identifier les sites de fixation de facteurs de transcription biologiquement actifs. La version automatisée de PromAn dédiée à l'analyse à haut débit facilitera la compréhension de réseaux de régulations complexes et surtout leurs impacts sur la santé et les maladies humaines.

Keywords/Mots clés: high-throughput promoter analysis/analyse de promoteurs à haut-débit, conservation of promoter regions in mammals/conservation des regions promotrices chez les mammifères, phylogenetic footprinting, PromAn, GOAnno, Gene Ontology.

Laboratory : Laboratoire de Bioinformatique et Génomique Intégratives (LBGI), UMR7104, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), 1 rue Laurent Fries, BP10142, 67404 Illkirch-Graffenstaden cedex.