ÉCOLE NATIONALE VÉTÉRINAIRE D'ALFORT

Année 2013

LES FACTEURS DE TRANSCRIPTION CHEZ LES MAMMIFÈRES, MALADIES ASSOCIÉES, MÉTHODES D'ÉTUDE ET APPLICATION DE LA PCR-SÉLECTION SUR STOX1

THÈSE

Pour le

DOCTORAT VÉTÉRINAIRE

Présentée et soutenue publiquement devant

LA FACULTÉ DE MÉDECINE DE CRÉTEIL

le……

par

Aurélien, Hervé DUCAT

Né le 29 mai 1988 à Laon (Aisne)

JURY

Président : Pr.
Professeur à la Faculté de Médecine de CRÉTEIL

Membres
Directeur : M^{me} Marie ABITBOL
Maître de conférences à l'ENVA
Assesseur : M. Laurent TIRET
Maître de conférences à l'ENVA
Invité : M^{me} Fanny PILOT-STORCK
Maître de conférences à l'ENVA

Mars 2013

LISTE DES MEMBRES DU CORPS ENSEIGNANT

Directeur : M. le Professeur GOGNY Marc

Directeurs honoraires : MM. les Professeurs : COTARD Jean-Pierre, MORAILLON Robert, PARODI André-Laurent, PILET Charles, TOMA Bernard

Professeurs honoraires : Mme et MM. : BENET Jean-Jacques, BRUGERE Henri, BRUGERE-PICOUX Jeanne, BUSSIERAS Jean, CERF Olivier, CLERC Bernard, CRESPEAU François, DEPUTTE Bertrand, MOUTHON Gilbert, MILHAUD Guy, POUCHELON Jean-Louis, ROZIER Jacques

DEPARTEMENT D'ELEVAGE ET DE PATHOLOGIE DES EQUIDES ET DES CARNIVORES (DEPEC)

Chef du département : M. POLACK Bruno, Maître de conférences - Adjoint : M. BLOT Stéphane, Professeur

UNITE DE CARDIOLOGIE
- Mme CHETBOUL Valérie, Professeur *
- Mme GKOUNI Vassiliki, Praticien hospitalier

UNITE DE CLINIQUE EQUINE
- M. AUDIGIE Fabrice, Professeur
- M. DENOIX Jean-Marie, Professeur
- Mme DUMAS Isabelle, Maître de conférences contractuel
- Mme GIRAUDET Aude, Praticien hospitalier *
- M. LECHARTIER Antoine, Maître de conférences contractuel
- Mme MESPOULHES-RIVIERE Céline, Praticien hospitalier
- Mme TRACHSEL Dagmar, Maître de conférences contractuel

UNITE D'IMAGERIE MEDICALE
- Mme BEDU-LEPERLIER Anne-Sophie, Maître de conférences contractuel
- Mme STAMBOULI Fouzia, Praticien hospitalier

UNITE DE MEDECINE
- Mme BENCHEKROUN Ghita, Maître de conférences contractuel
- M. BLOT Stéphane, Professeur*
- Mme MAUREY-GUENEC Christelle, Maître de conférences

UNITE DE MEDECINE DE L'ELEVAGE ET DU SPORT
- Mme CLERO Delphine, Maître de conférences contractuel
- M. GRANDJEAN Dominique, Professeur *
- Mme YAGUIYAN-COLLIARD Laurence, Maître de conférences contractuel

DISCIPLINE : NUTRITION-ALIMENTATION
- M. PARAGON Bernard, Professeur

DISCIPLINE : OPHTALMOLOGIE
- Mme CHAHORY Sabine, Maître de conférences

UNITE DE PARASITOLOGIE ET MALADIES PARASITAIRES
- M. BENSIGNOR Emmanuel, Professeur contractuel
- M. BLAGA Radu Gheorghe, Maître de conférences (rattaché au DPASP)
- M. CHERMETTE René, Professeur *
- M. GUILLOT Jacques, Professeur
- Mme MARIGNAC Geneviève, Maître de conférences
- M. POLACK Bruno, Maître de conférences

UNITE DE PATHOLOGIE CHIRURGICALE
- M. FAYOLLE Pascal, Professeur
- M. MAILHAC Jean-Marie, Maître de conférences
- M. MOISSONNIER Pierre, Professeur*
- M. NIEBAUER Gert, Professeur contractuel
- Mme RAVARY-PLUMIOEN Bérangère, Maître de conférences (rattachée au DPASP)
- Mme VIATEAU-DUVAL Véronique, Professeur
- M. ZILBERSTEIN Luca, Maître de conférences

DISCIPLINE : URGENCE SOINS INTENSIFS
- Vacant

DEPARTEMENT DES PRODUCTIONS ANIMALES ET DE LA SANTE PUBLIQUE (DPASP)

Chef du département : M. MILLEMANN Yves, Professeur - Adjoint : Mme DUFOUR Barbara, Professeur

UNITE D'HYGIENE ET INDUSTRIE DES ALIMENTS D'ORIGINE ANIMALE
- M. AUGUSTIN Jean-Christophe, Maître de conférences
- M. BOLNOT François, Maître de conférences *
- M. CARLIER Vincent, Professeur
- Mme COLMIN Catherine, Maître de conférences

UNITE DES MALADIES CONTAGIEUSES
- Mme DUFOUR Barbara, Professeur*
- Mme HADDAD/HOANG-XUAN Nadia, Professeur
- Mme PRAUD Anne, Maître de conférences
- Mme RIVIERE Julie, Maître de conférences contractuel

UNITE DE PATHOLOGIE MEDICALE DU BETAIL ET DES ANIMAUX DE BASSE-COUR
- M. ADJOU Karim, Maître de conférences *
- M. BELBIS Guillaume, Assistant d'enseignement et de recherche contractuel
- M. HESKIA Bernard, Professeur contractuel
- M. MILLEMANN Yves, Professeur

UNITE DE REPRODUCTION ANIMALE
- Mme CONSTANT Fabienne, Maître de conférences
- M. DESBOIS Christophe, Maître de conférences (rattaché au DEPEC)
- M. FONTBONNE Alain, Maître de conférences (rattaché au DEPEC)
- Mme MASSE-MOREL Gaëlle, Maître de conférences contractuel
- M. MAUFFRE Vincent, Assistant d'enseignement et de recherche contractuel
- M. NUDELMANN Nicolas, Maître de conférences (rattaché au DEPEC)
- M. REMY Dominique, Maître de conférences*

UNITE DE ZOOTECHNIE, ECONOMIE RURALE
- M. ARNE Pascal, Maître de conférences*
- M. BOSSE Philippe, Professeur
- M. COURREAU Jean-François, Professeur
- Mme GRIMARD-BALLIF Bénédicte, Professeur
- Mme LEROY-BARASSIN Isabelle, Maître de conférences
- M. PONTER Andrew, Professeur

DEPARTEMENT DES SCIENCES BIOLOGIQUES ET PHARMACEUTIQUES (DSBP)

Chef du département : Mme COMBRISSON Hélène, Professeur - Adjoint : Mme LE PODER Sophie, Maître de conférences

UNITE D'ANATOMIE DES ANIMAUX DOMESTIQUES
- M. CHATEAU Henry, Maître de conférences*
- Mme CREVIER-DENOIX Nathalie, Professeur
- M. DEGUEURCE Christophe, Professeur
- Mme ROBERT Céline, Maître de conférences

DISCIPLINE : ANGLAIS
- Mme CONAN Muriel, Professeur certifié

UNITE DE BIOCHIMIE
- M. BELLIER Sylvain, Maître de conférences*
- M. MICHAUX Jean-Michel, Maître de conférences

DISCIPLINE : BIOSTATISTIQUES
- M. DESQUILBET Loïc, Maître de conférences

DISCIPLINE : EDUCATION PHYSIQUE ET SPORTIVE
- M. PHILIPS Pascal, Professeur certifié

DISCIPLINE : ETHOLOGIE
- Mme GILBERT Caroline, Maître de conférences

UNITE DE GENETIQUE MEDICALE ET MOLECULAIRE
- Mme ABITBOL Marie, Maître de conférences
- M. PANTHIER Jean-Jacques, Professeur*

UNITE D'HISTOLOGIE, ANATOMIE PATHOLOGIQUE
- Mme CORDONNIER-LEFORT Nathalie, Maître de conférences*
- M. FONTAINE Jean-Jacques, Professeur
- Mme LALOY Ève, Maître de conférences contractuel
- M. REYES GOMEZ Édouard, Assistant d'enseignement et de recherche contractuel

UNITE DE PATHOLOGIE GENERALE MICROBIOLOGIE, IMMUNOLOGIE
- M. BOULOUIS Henri-Jean, Professeur
- Mme LE ROUX Delphine, Maître de conférences
- Mme QUINTIN-COLONNA Françoise, Professeur*

UNITE DE PHARMACIE ET TOXICOLOGIE
- Mme ENRIQUEZ Brigitte, Professeur
- M. PERROT Sébastien, Maître de conférences
- M. TISSIER Renaud, Professeur*

UNITE DE PHYSIOLOGIE ET THERAPEUTIQUE
- Mme COMBRISSON Hélène, Professeur
- Mme PILOT-STORCK Fanny, Maître de conférences
- M. TIRET Laurent, Maître de conférences*

UNITE DE VIROLOGIE
- M. ELOIT Marc, Professeur
- Mme LE PODER Sophie, Maître de conférences *

*** responsable d'unité**

REMERCIEMENTS

Au Professeur de la faculté de Médecine de Créteil,
Pr.
Qui m'a fait l'honneur d'accepter la présidence de mon jury de thèse,
Hommage respectueux.

À ma directrice de thèse,
Madame le Docteur Marie Abitbol,
Maître de conférences à l'École Nationale Vétérinaire d'Alfort,
Pour votre immense gentillesse, votre grande disponibilité, vos nombreux conseils tant sur le plan scientifique que professionnel ou personnel et votre soutien dans mon choix de la voie de la recherche. Vous m'avez littéralement transmis votre passion pour la génétique et la science fondamentale lors de vos cours, et vous resterez pour moi l'une des plus grandes enseignantes pédagogues que j'ai connues.
Sincères remerciements.

À mon assesseur de thèse,
Monsieur le Docteur Laurent Tiret,
Maître de conférences à l'École Nationale Vétérinaire d'Alfort,
Pour avoir bien voulu juger deux fois mon travail après la lecture de mon rapport de stage de M2, et pour m'avoir transmis le goût des sciences de la reproduction et de la recherche lors de vos cours.
Sincères remerciements.

À mon invitée,
Madame le Docteur Fanny Pilot-Storck,
Maître de conférences à l'École Nationale Vétérinaire d'Alfort,
Pour votre immense gentillesse, votre grande disponibilité, vos nombreux conseils tant sur le plan scientifique que professionnel ou personnel et votre soutien dans mon choix de la voie de la recherche. Vous entretenez un rapport avec les étudiants que je trouve sensationnel et pour couronner le tout vous êtes dotée d'un sens de l'humour incroyable qui va jusqu'à la participation à l'une des vidéos du spectacle de l'Accueil 2013 !
Sincères remerciements.

À mon directeur de stage de Recherche,
Monsieur le Docteur Daniel Vaiman,
Responsable de l'équipe « Génomique, Épigénétique et Physiopathologie de la Reproduction » à l'Institut Cochin (INSERM U1016 CNRS UMR8104),
Pour m'avoir accueilli au sein du laboratoire, pour tes conseils et ton soutien dans mon projet de réaliser une thèse de science, pour ton humour et pour toutes tes connaissances infinies sur l'évolution et la diversité des espèces animales que tu fais partager avec tellement de passion. Je retiens de toi pour l'instant deux citations philosophiques de la recherche : « Rien n'a de sens en biologie, si ce n'est à la lumière de l'évolution » (Théodosius Dobzhansky) et « Ce n'est pas en améliorant la bougie qu'on a inventé l'électricité » (Niels Henrik David Bohr (1885-1962), physicien danois, prix Nobel de physique de 1922 pour son apport à l'édification de la mécanique quantique). Essayons de les appliquer au mieux pour faire avancer la science ne serait-ce qu'un peu au cours de ces trois années de travail collaboratif, sans se laisser happer par nos obligations administratives respectives…
Sincères remerciements.

À mon encadrante de stage de M2,
Madame le Docteur Ludivine Doridot,
Merci est un bien faible mot pour exprimer toute la patience, le sérieux, le professionnalisme et la rigueur dont tu as fait preuve pendant toute cette année d'encadrement : tu m'as formé aux techniques de laboratoire nécessaires à l'élaboration de mon projet, tu m'as suivi et supporté dans toutes mes manips, dans ma demande de bourse de thèse et jusque dans la rédaction de mon rapport de stage. Tu mérites le meilleur… Bon courage pour ton post-doc aux États-Unis !
Mes plus sincères remerciements.

À toute l'équipe du laboratoire « Génomique, Épigénétique et Physiopathologie de la Reproduction » de l'Institut Cochin,
À Céline Méhats, pour ton généreux soutien scientifique et technique dans mon projet (demande de bourse, lecture de mon rapport de stage) et pour toutes les discussions scientifiques enrichissantes que nous avons pu avoir ensemble (et qui j'espère continueront).
Merci aussi à tout le reste de l'équipe : Jana (pour m'avoir appris tous les secrets du Western blot), Rosa, Isabelle, Ami (qui m'a initié aux dons pour la science !), Ahmed, Sandrine, Laurence, Come, Patrick, Julie, Brigitte,… Merci également à tous les étudiants M2 que j'ai pu rencontrer cette année et avec qui j'ai passé d'excellents moments au labo : Charlotte (ma désormais co-thésarde), Adrianne, Lucile, Sandrine, Aurélie, Adrien, Laetitia, Laila.
À tous, merci d'avoir si bien accueilli un étudiant véto comme moi, assez fou pour se lancer dans la voie de la recherche… Merci à tous de participer à cette excellente ambiance dans le labo, où les journées sont rythmées de pauses café, de lectures d'horoscope, de secrets de la calculette de l'amour et de discussions aussi bien scientifiques que culturelles. Grâce à vous je me sens tellement bien dans ce labo et dans ce que je fais que j'ai eu envie d'y rester trois ans pour une thèse !
À tous mes très chaleureux remerciements.

Aux plates-formes de l'Institut Cochin,
Avec qui j'ai initié des collaborations. Pour votre accueil et votre rapidité,
Merci.

À ma famille de sang, mes amis de toujours,
À la famille alforienne et à la grande famille vétérinaire, mes deux nouvelles familles de cœur depuis cinq ans,
À toutes les belles personnes que j'ai rencontrées, trop nombreuses pour être citées ici mais qui figurent dans mes remerciements personnels plus développés. Merci d'être toujours là pour me soutenir dans mes projets professionnels ou personnels.

À Claude Bourgelat (1712-1779), créateur des écoles vétérinaires, fondateur de la profession vétérinaire et promoteur du concept de biopathologie comparée, sans lequel la médecine moderne n'aurait jamais connus aussi rapidement les fantastiques progrès qu'elle a connu au cours des deux derniers siècles.
Aujourd'hui, je fais le serment d'avoir à tout moment et en tout lieu le souci constant de la dignité et de l'honneur de la profession vétérinaire, que je m'efforcerai de transmettre au mieux en hommage à Claude Bourgelat :
« La fortune consiste moins dans le bien que l'on a que dans celui que l'on peut faire » (*Règlement pour les Écoles royales Vétérinaires*, 1777).

Gravure de Pigeot, Lyon, 2ème moitié du 18ème siècle. Inv. 37.289

TABLE DES MATIÈRES

LISTE DES FIGURES

*

LISTE DES ABRÉVIATIONS UTILISÉES

ADN : Acide Désoxyribonucléique
ADNc : Acide Désoxyribonucléique complémentaire
AMPc : Adénosine monophosphate cyclique
AP1 : *Activator Protein 1*, facteur de transcription formé par l'hétérodimérisation de protéines des familles c-Jun et c-Fos impliquées dans la régulation du cycle cellulaire (oncoprotéines).
ARN : Acide Ribonucléique
ARNm : ARN messager
ARNnc : ARN de non-codage ou non-codant (ARN ne codant pas pour une protéine)
ARNr : ARN ribosomal
ARNt : ARN de transfert
ATP : Adénosine triphosphate
Bdp1 : *B double prime 1*
Brf1 : *TFIIIB-related factor 1*
BSA : *Bovine Serum Albumin* (Albumine de Sérum Bovin)
CAK : *Cyclin-Activating Kinase*
CBP/p300 : *CREB Binding Protein*
ChIP : *Chromatin ImmunoPrecipitation*
c-Jun : proto-oncogène découvert par une équipe japonaise à partir du virus ASV17 (*avian sarcoma virus 17*) d'où son nom (*jû nana* signifie 17 en Japonais)
c-Fos : *Finkel-Biskis-Jenkins murine sarcoma virus (FBJ MSV) oncogene homolog*, proto-oncogène
CMV minimal (Cytomégalovirus) : séquence promotrice forte permettant l'initiation de la transcription (contenant notamment la boîte TATA, la boîte CAAT,…).
c-Myc : *v-myc avian myelocytomatosis viral oncogene homolog*, facteur de transcription surexprimé dans de nombreux cancers humains
CNV : *Copy Number Variation*
CPE : *Core Promoter Element*
CRE : *Cyclic-AMP Response Element* (élément de réponse à l'AMP cyclique)
CREB : *Cyclic-AMP Response Element-binding protein*
CTF : *CAAT box transcription factor*
Da : Dalton (unité de mesure standard, utilisée pour mesurer la masse des atomes et des molécules)
DBD : *DNA Binding Domain*
DES : *Distal Sequence Element*
EBNA : *Epstein-Barr Nuclear Antigen*
EDTA : *Ethylenediaminetetraacetic acid* (Acide Éthylène Diamine Tétracétique)
ENCODE : *Encyclopedia Of DNA Elements*
ER : *Estrogen Receptor* (récepteur à l'œstrogène)
ERE : *Estrogen Responsive Element* ou *Estrogen Response Element*
FOX : *Forkhead box*
FRET : *Fluorescence Resonance Energy Transfert*
GFP : *Green Fluorescent Protein*
GRE : *Glucocorticoid Responsive Element*
GST : Glutathion S-transférase
GTF : *General Transcription Factors*, facteurs généraux de la transcription dans une cellule eucaryote
HAT : histone acétyltransférase
HDAC : histone désacétylase
HLA : *Human Leucocyte Antigen*

HRP : *Horseradish peroxidase* (peroxidase de raifort)
HSF : *heat shock transcription factors*
ICR : *Internal Control Région*
IDR : *Intrinsically Disordered Regions*
IKB : *Inhibitor of κB*
iPSCs : *induced Pluripotent Stem Cells* (cellules souches pluripotentes induites)
IPTG : Isopropyl-bêta-D-1-thiogalactopyranoside, analogue de l'allolactose, il se lie au répresseur de l'opéron lactose et bloque son action, ce qui induit en particulier la transcription du gène de la bêta galactosidase (ou *lacZ*).
kb : kilobase
kDa : kiloDalton
KI : *Knock-in*
KID : *Kinase Inductible Domain*
KO : *Knock-out*
LB : *Lysogeny Broth* (littéralement bouillon lysogène) ou incorrectement Milieu Luria-Bertani, milieu de culture nutritif servant initialement à la culture bactérienne.
LINE : *Long Interspersed Nuclear Element*
LTR : *Long Terminal Repeat*
M : mol/L
Max : *MYC associated factor X*, facteur de transcription possédant un motif de type fermeture éclair à leucines
Mb : Mégabase
MEME : *Multiple Em for Motif Elicitation*
MG132 (ou Z-Leu-Leu-Leu-al) : Inhibiteur du protéasome
MHC : *Major Histocompatibility Complex* (complexe majeur d'histocompatibilité)
MMTV : *Mouse Mammary Tumor Virus*
NCBI : *National Centre for Biotechnology Information*
NES : *Nuclear Export Signal*
NF1 : *Nuclear factor 1*, facteur se liant à la boîte CAAT
NF-κB : *Nuclear Factor kappa-light-chain-enhancer of activated B cells*, facteur de transcription majeur dans la transmission des signaux impliqués dans l'inflammation
NHGRI : *National Human Genome Research Institute*
NK : *Natural Killer*
NLS : *Nuclear Localization Sequence*
NP-40 : Tergitol-type NP-40 (*Nonyl Phenoxypolyethoxylethanol*)
Oct1 : *Octamer transcription factor 1*
OMIA : *Online Mendelian Inheritance in Animals*
PAS : Pression Artérielle Systolique
PAD : Pression Artérielle Diastolique
pb : paire de bases
PBS : *Phosphate Buffer Saline*
PCR : *Polymerase Chain Reaction* (réaction de polymérisation en chaîne)
PE : prééclampsie
PKA : Protéine Kinase A
PM : Poids Moléculaire
Pol I, II ou III : ARN polymérase I, II ou III (enzyme synthétisant une séquence d'ARN à partir d'une séquence ADN)
PSE : *Proximal Sequence Element*
P-TEFb : *Positive Transcription Elongation Factor b*
PTF : *PSE-binding transcription factor*, aussi appelé SNAPc
PVDF : *PolyVinyliDene Fluoride* (polyfluorure de vinylidène)
RE : *Responsive Element*

RIPA : *Radioimmunoprecipitation Assay*, tampon d'extraction protéique à partir de culture cellulaire.
rpm : *revolutions per minute* (tours par minute)
SAP : *Shrimp Alkaline Phosphatase*
SL1 : *Selectivity Factor 1*
SDS : *Sodium Dodecyl Sulfate* (DodécylSulfate de Sodium ou laurylsulfate de sodium)
SDS-PAGE : *Sodium Dodecyl Sulfate-Polyacrylamide gel electrophoresis* (électrophorèse sur un gel SDS-polyacrylamide)
sENG : Endogline soluble
SF1 : *Steroidogenic Factor 1*.
sFLT-1 : *soluble Fms-like tyrosine kinase 1*, récepteur soluble du VEGF
snRNAs : *small nuclear RNA*, petits ARN nucléaires (pARNn) impliqués dans l'épissage des ARNm dans les cellules eucaryotes.
SNAPc : *snRNA activating protein complex*, aussi appelé PTF
snoRNAs : *small nucleolar RNAs*, petits ARN nucléolaires qui aident à la maturation des ARNr dans les cellules eucaryotes.
SNP : *Single Nucleotide Polymorphism* (polymorphisme d'une seule base)
Sp1 : *Specificity protein 1*, protéine en doigt de zinc qui reconnaît spécifiquement la séquence GGGCGG (parfois appelée GC *box*), séquence promotrice de nombreux gènes de ménage
SOB : *Super Optimal Broth*, milieu de culture riche en nutriments (proche du LB) utilisé pour la culture microbiologique, généralement d'*Escherichia coli*.
SOC : *Super Optimal broth with Catabolite repression* (SOB avec du glucose ajouté). La culture d'*Escherichia coli* dans du SOB ou du SOC rend les transformations de plasmides plus efficaces.
STOX1 : *Storkhead box 1*
STRE : *STOX1 Responsive Element*
SVF : sérum de veau fœtal
TAF : *TBP-Associated Factors*
TBP : *TATA Binding Protein*
TF : *Transcription Factor*
TGF-β : *Transforming Growth Factor beta*
TIC : *TAF and Initiator-dependent Cofactors*
Tris : *Trishydroxyméthylaminométhane*
UBF1 : *Upstream-binding factor 1*
UCE : *Upstream Control Element*
UV : UltraViolet
VEGF : *Vascular endothelial growth factor* (facteur de croissance de l'endothélium vasculaire)
WB : *Western blot*
WT : *Wild Type* (sauvage)
X-gal : *5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside* ($C_{14}H_{15}BrClNO_6$) : hétéroside du galactose lié à un noyau indole substitué, substrat artificiel de la bêta-galactosidase.
XPB : *Xeroderma Pigmentosum B*
XPD : *Xeroderma Pigmentosum D*

INTRODUCTION

Les facteurs de transcription sont classiquement définis comme des protéines ayant la capacité de se lier à des séquences d'ADN spécifiques et de réguler la transcription (R. Hughes, 2011). Les facteurs de transcription ont longtemps fasciné les biologistes moléculaires car ils ont souvent été identifiés, au moins au départ, comme des clés de la compréhension de la régulation complexe de l'expression génique. Ils interviennent notamment dans le métabolisme cellulaire, l'homéostasie, et dans la différenciation des cellules au cours du développement embryonnaire. Les premières applications utilisant ces protéines commencent à apparaître et sont prometteuses, mais leurs actions diverses et variées et leurs mécanismes d'action non complètement élucidés complexifient les mises au point. De par leur importance dans les processus biologiques, une mutation dans un facteur de transcription risque souvent d'entraîner des répercussions importantes sur l'organisme, pouvant même être létales à court ou à long terme. Ainsi, plusieurs phénotypes et maladies héréditaires chez les mammifères trouvent leur origine dans des mutations de facteurs de transcription.

Dans une première partie bibliographique, je présenterai une synthèse des données actuelles concernant les facteurs de transcription, puis j'aborderai des exemples de phénotypes et maladies héréditaires dues à des mutations dans des facteurs de transcription chez les mammifères, avant de m'intéresser aux méthodes de biologie moléculaire qui permettent l'étude des facteurs de transcription.

Dans une seconde partie, je présenterai les résultats expérimentaux que j'ai obtenus au cours du stage que j'ai réalisé dans le laboratoire du docteur Daniel Vaiman à l'Institut Cochin. J'ai travaillé sur une maladie humaine de la grossesse : la prééclampsie. C'est une maladie qui survient chez 3 à 8 % des femmes enceintes et qui est définie par une hypertension gestationnelle (Pression Artérielle Systolique (PAS) $\geq$ 140 mmHg et Pression Artérielle Diastolique (PAD) $\geq$ 90 mmHg) et une protéinurie (> 300 mg/24 h) se développant à partir de la 20ème semaine d'aménorrhée. La maladie peut s'aggraver jusqu'à entraîner le décès de la mère (environ 20 décès par an en France). À ce jour, les traitements préventifs et thérapeutiques sont très limités, et cela est principalement dû au fait que la physiopathologie de cette maladie est complexe et n'est actuellement pas totalement élucidée. Il a été montré toutefois que cette maladie possède une forte composante génétique : plusieurs gènes y ont été associés, notamment *STOX1 (Storkhead box 1)*, qui a été mis en évidence grâce à l'étude de familles hollandaises. Par analyse bio-informatique et alignement avec plusieurs facteurs de transcription, il a été prédit que ce gène code un facteur de transcription. Cependant, on ne connaît ni sur quelle(s) séquence(s) il se fixe, ni quels potentiels gènes il régule. Mon travail a consisté à identifier le site de fixation à l'ADN de la protéine STOX1 en essayant de déterminer une séquence consensus via la méthode de PCR-sélection. Le principe de cette technique est de purifier des oligonucléotides sur lesquels un facteur transcriptionnel se fixe, au sein d'une banque d'oligonucléotides aléatoires. La première partie de mon stage a consisté à construire un vecteur plasmidique permettant l'expression d'une protéine chimérique Flag-STOX1, ce qui m'a permis d'utiliser un anticorps anti-Flag très spécifique lors de la phase de purification. Ensuite, j'ai produit cette protéine chimérique en grande quantité par transfection du plasmide d'expression dans des cellules COS-7. J'ai contrôlé le succès de cette transfection par Western blot, puis ai procédé aux étapes de PCR-sélection. Les oligonucléotides séquencés ont été analysés afin de déterminer un élément consensus de fixation. L'analyse bio-informatique réalisée grâce au logiciel MEME m'a permis de mettre en évidence deux séquences consensus majoritaires. J'ai commencé à valider ces séquences par deux méthodes : i) par du retard sur gel afin de valider l'interaction physique ; ii) en intégrant les éléments identifiés devant le gène rapporteur luciférase afin de valider l'effet fonctionnel. Je discuterai des résultats que j'ai obtenus avant de conclure.

PREMIÈRE PARTIE : ANALYSE BIBLIOGRAPHIQUE

LES FACTEURS DE TRANSCRIPTION CHEZ LES MAMMIFÈRES

I. Qu'est-ce qu'un facteur de transcription ?

A. Introduction

D'après (Griffiths *et al.*, 2013).

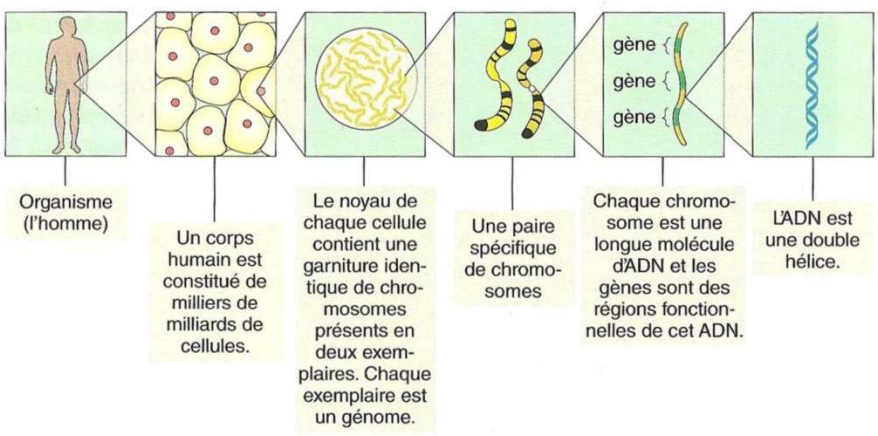
Les mammifères sont des organismes pluricellulaires eucaryotes, c'est-à-dire constitués d'une multitude de cellules possédant un noyau contenant l'ADN (acide désoxyribonucléique), molécule essentielle à la vie (Figure 1). Cette année, nous fêtons les 60 ans de la découverte par Watson et Crick de la structure en double hélice de l'ADN (Figure 2), qui a révolutionné la science et a eu un impact majeur en biologie et notamment en génétique, alors qu'à l'époque aucun scientifique ne soupçonnait l'importance d'une telle découverte. En effet, un gène était auparavant défini comme un caractère héréditaire (facteur mendélien) de nature inconnue. Depuis, il a été mis en évidence que l'information biologique (c'est-à-dire ce qui est nécessaire pour créer la forme biologique) est codée par la molécule d'ADN, divisée en unités fonctionnelles appelées gènes. La forme biologique est en grande partie un produit des protéines de l'organisme. La génétique moléculaire a montré que la plupart des gènes codent une protéine spécifique, qui est synthétisée en deux étapes : au cours de l'étape de transcription, l'ARN (acide ribonucléique) est transcrit à partir de l'ADN, et au cours de l'étape de traduction, l'ARN est « lu » pour synthétiser la séquence d'acides aminés de la protéine (Figure 3).

La première étape du transfert de l'information du gène à la protéine est donc la synthèse d'un brin d'ARN dont la séquence de bases correspond à celle du fragment d'ADN du gène. C'est sur ce processus (qui a été appelé transcription car il rappelle la transcription ou copie de mots) que va se focaliser mon propos. Chaque portion d'ADN transcrite est encadrée (flanquée) par une ou plusieurs régions qui déterminent à quels moments et dans quelles cellules aura lieu la transcription du gène. La région 5' non transcrite du gène notamment, a été appelée région promotrice (par analogie avec les systèmes procaryotes) et nous verrons qu'elle a une grande importance : c'est à ce niveau que va se fixer entre autres l'ARN polymérase, enzyme qui va synthétiser l'ARN à partir de l'ADN. L'unité transcriptionnelle globale composée d'une région codant l'ARN et de ses éléments régulateurs flanquants est l'unité fonctionnelle élémentaire du génome que les scientifiques appellent désormais gène.

La transcription est classiquement décrite en trois étapes : initiation, élongation et terminaison (même si la séparation est en réalité artificielle car, comme tout processus biologique, il s'agit plus d'un continuum et non pas d'étapes indépendantes). L'événement primordial dans la transcription est l'initiation, phénomène complexe chez les eucaryotes, qui met en jeu trois types d'acteurs : des facteurs de transcription, l'ARN polymérase, et des facteurs médiateurs (*mediators*). C'est à cette étape que je vais m'intéresser dans cette thèse, et en particulier à l'étude des facteurs de transcription.

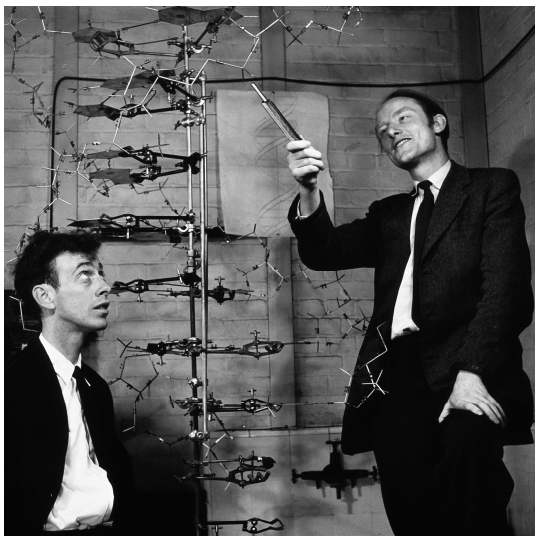
Un facteur de transcription, ou facteur transcriptionnel, est une protéine (parfois appelée protéine *trans*-régulatrice) interagissant avec une séquence spécifique d'ADN située en amont d'un gène appelé gène-cible, et modifiant le taux de transcription de celui-ci (Kaplan et Delpech, 2007). Nous verrons ainsi que la transcription d'un gène est gouvernée d'une part par des séquences promotrices, sur lesquelles se fixe l'ARN polymérase et les facteurs généraux de la transcription (GTF ou *General Transcription Factors*, qui font partie des facteurs de transcription), et d'autre part par des séquences régulatrices en amont, sur lesquelles se fixent spécifiquement d'autres facteurs de transcription, qui peuvent aussi bien être activateurs qu'inhibiteurs de ladite transcription.

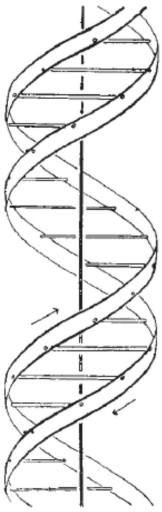
Figure 1 : Chaque cellule d'un organisme contient un matériel génétique sous forme d'ADN

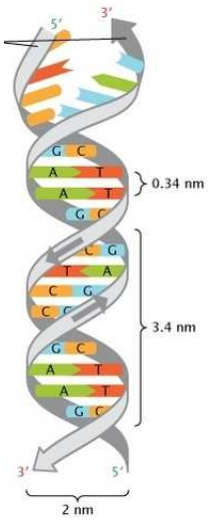


D'après (Griffiths *et al.*, 2013). Les 100 000 milliards de cellules qui composent un être humain possèdent chacune la même information génétique, sous forme d'ADN.

Figure 2 : La découverte de la structure de l'ADN en 1953 par Watson et Crick

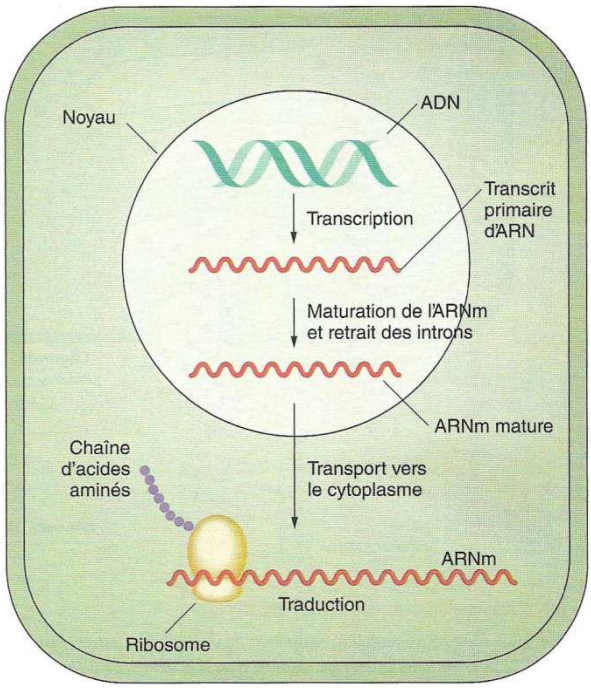
A. 

B. 

C. 

Sur la photographie (A, *Science Photo Library*), sans doute la plus célèbre du monde scientifique du 20ème siècle, on peut voir James Dewey Watson (né en 1928), assis à gauche, et Francis Harry Compton Crick (1916-2004), debout à droite, admirant leur modèle de double hélice de la structure de l'ADN (photographie prise par M. Anthony Barrington Brown, en mai 1953 au laboratoire de Cavendish à l'Université de Cambridge au Royaume-Uni). Le travail de Crick et Watson sur la structure de l'ADN, depuis leur rencontre en 1951, a été réalisé avec la connaissance des ratios des bases dans l'ADN déterminés par Chargaff, et l'accès à quelques données de cristallographie aux rayons X réalisées par Maurice Hugh Frederick Wilkins et Rosalind Elsie Franklin au *King's College* de Londres. La combinaison de tous ces travaux a mené à la déduction de la structure en double hélice de l'ADN (structure à deux chaînes hélicoïdales enroulées chacune autour du même axe). Cette année, nous fêtons les 60 ans de la publication dans le 171ème volume de la revue scientifique britannique *Nature* (daté du 25 avril 1953) des trois premières publications décrivant cette découverte (Franklin et Gosling, 1953 ; Watson et Crick, 1953 ; Wilkins *et al.*, 1953). Le schéma B est celui qui figure dans le papier princeps de Watson et Crick, accompagné de cette modeste légende : « cette figure est purement schématique. Les deux rubans symbolisent les deux chaînes de sucre-phosphate, et les tiges horizontales les paires de bases maintenant les chaînes ensemble. La ligne verticale marque l'axe de la fibre. » (Watson et Crick, 1953). La figure C est une représentation actuelle de la molécule d'ADN (Pray, 2008). Le prix Nobel de Physiologie ou Médecine en 1962 a été décerné conjointement à Crick, Watson et Wilkins (Franklin étant morte d'un cancer en 1958), « pour leurs découvertes concernant la structure moléculaire des acides nucléiques et leur signification pour le transfert de l'information dans le matériel vivant ».

Figure 3 : La transcription et la traduction dans une cellule eucaryote



D'après (Griffiths *et al.*, 2013). Selon le schéma général, dans une cellule eucaryote, l'ARNm (ARN messager) est transcrit à partir de l'ADN présent dans le noyau, puis subit chez la plupart des eucaryotes une série de modifications (excision des introns notamment) et est transporté vers le cytoplasme qui contient toute la machinerie nécessaire pour finalement traduire l'ARNm en une chaîne polypeptidique. La forme initiale des transcrits des gènes destinés à la synthèse protéique s'appelle l'ARN messager (ARNm). Le terme messager est utilisé pour souligner l'idée que cette molécule est le véhicule qui transporte l'information d'un gène jusqu'à la machinerie de synthèse protéique. Ce processus, qui n'était qu'une théorie au départ, a été nommé « *central dogma of molecular biology* » (« dogme central de la biologie moléculaire ») par Francis Crick (Crick, 1970). Le mot dogme prête ici à confusion, car il s'agit plutôt d'une hypothèse scientifique et non pas d'une doctrine établie comme une vérité incontestable. En anglais, le terme *dogma* renvoie également à « une idée qui n'est pas étayée par des preuves rationnelles », ce qui à l'époque était exact. La formulation fut malheureusement longtemps restée ainsi alors que la théorie en elle-même a été plusieurs fois remise en cause. Aujourd'hui les biologistes préfèrent utiliser l'expression « théorie fondamentale de la biologie moléculaire » pour désigner le modèle schématique de la conservation et de l'utilisation de l'information génétique qui se résume ainsi : « l'ADN dirige sa propre réplication en ADN identique, ainsi que sa transcription en ARN, pouvant ou non être traduit en protéines ».

B. L'initiation de la transcription chez les eucaryotes

D'après (Kaplan et Delpech, 2007).

Chez les eucaryotes, les mécanismes de transcription mettent en jeu 3 types d'ARN polymérases, chacune spécifique d'une classe de gènes : l'ARN polymérase I (Pol I) transcrit les ARN des ribosomes (ARNr 28 S, 18 S et 5,8 S), l'ARN polymérase III (Pol III) transcrit les petits ARN (ARN de transfert ou ARNt, ARN 5S, d'autres ARN particuliers…) et l'ARN polymérase II (Pol II) transcrit essentiellement les ARN messagers (codant des polypeptides), mais aussi partiellement certains ARN non-codants ou ARNnc (certains *small nuclear RNA* ou *snRNA*, impliqués dans l'épissage des ARNm, et certains *snoRNA* ou *small nucleolar RNA*, petits ARN présents dans le nucléole qui aident à la maturation des ARNr). D'une manière générale, les ARN polymérases ne peuvent initier par elles-mêmes la transcription. La combinaison de courtes séquences dans le voisinage immédiat du gène constitue un signal de reconnaissance permettant à plusieurs facteurs de se lier à l'ADN afin de guider et d'activer la polymérase. Un ensemble essentiel de ces courtes séquences est souvent regroupé en amont de la séquence codante du gène et constitue le promoteur (voir plus loin). Je décrirai surtout les mécanismes mettant en jeu la Pol II, ceux des deux autres polymérases étant similaires.

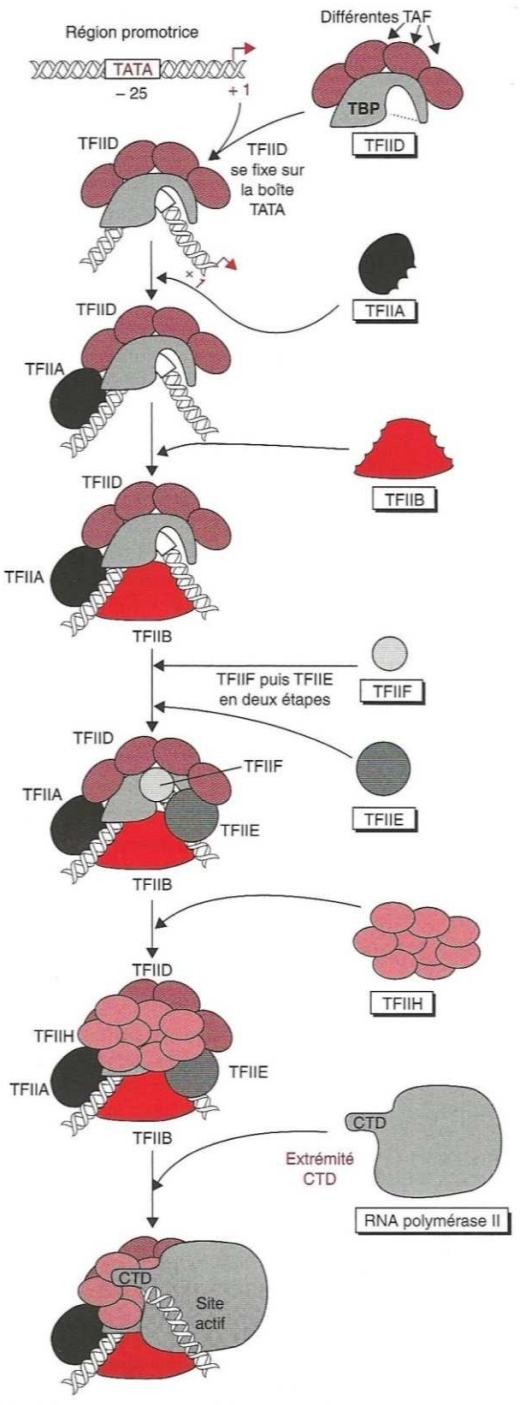
1. L'initiation de la transcription avec l'ARN polymérase II

La structure de l'ARN polymérase II et son mécanisme d'action sont longtemps restés obscurs, principalement parce qu'elle était purifiée de concert avec toute une série de protéines dont il n'était pas possible de déterminer s'il s'agissait de sous-unités ou de protéines contaminantes. L'ARN polymérase II est une enzyme complexe constituée de 12 sous-unités (Rpb1 à Rpb12) chez les mammifères. Les deux plus grosses sous-unités, Rpb1 et Rpb2, portent le site actif, les autres permettent l'interaction avec l'ADN, la stabilisation, des interactions avec des facteurs transcriptionnels ou une forte processivité par exemple (une enzyme processive est une enzyme capable de catalyser la même réaction de façon répétée le long d'un polymère sans se détacher de la chaîne, définition d'après (Alberts *et al.*, 2011)). À l'extrémité carboxy-terminale de la plus grande sous-unité Rpb1 se trouve une structure spécifique remarquable, qui n'est retrouvée dans aucune autre polymérase, appelée CTD (*Carboxy Terminal Domain*), et qui est constituée d'une répétition du motif Tyr-Ser-Pro-Thr-Ser-Pro-Ser (52 fois chez les mammifères). Des études de délétion ont pu montrer que cette structure était indispensable et jouait un rôle important dans la transition entre l'initiation et l'élongation.

L'ARN polymérase II ne se fixe pas directement sur l'ADN, mais par l'intermédiaire de plusieurs facteurs. Les premiers à intervenir sont les facteurs généraux de transcription (GTF, *General Transcription Factors*). Le mécanisme décrit dans la Figure 4 est séquentiel et correspond à ce qui doit être réalisé *in vitro* pour obtenir une transcription. Il n'est pas certain qu'il en soit ainsi *in vivo* et il est possible que la cellule utilise des complexes préformés.

La transcription nécessite aussi des acteurs complémentaires que l'on appelle médiateurs (*mediators*). Ce sont des protéines qui participent aux mécanismes de l'initiation et de l'élongation de la transcription et qui ont une action activatrice pour certaines et inhibitrice pour d'autres. Ces médiateurs s'associent entre eux et avec les protéines déjà décrites en formant des complexes de composition très variable, qui peuvent contenir jusqu'à 18 protéines. Les mécanismes impliqués ne sont pas encore complètement connus. Ils ont été d'abord caractérisés chez la levure, qui en possède une vingtaine. Chez l'homme, la famille des médiateurs est constituée d'au moins huit membres (dont certaines kinases).

Figure 4 : Mécanisme d'initiation de la transcription avec l'ARN Pol II



D'après (Kaplan et Delpech, 2007). Le premier acteur est le facteur TFIID (*transcription factor* IID) qui a une structure particulièrement complexe. Sa composition est variable en fonction des tissus et de l'état cellulaire. On y retrouve toujours une sous-unité appelée TBP (*TATA Binding Protein*) et une dizaine de polypeptides appelés TAF (*TBP-Associated Factors*). TBP, qui a une forme qui ressemble à une selle, se fixe sur une séquence riche en T et en A, la boîte TATA (*TATA box*[1]), située 25 à 30 bases en amont du premier nucléotide qui sera transcrit. TBP[2] interagit avec la boîte TATA au niveau du petit sillon de l'ADN et induit une torsion d'environ 80 ° de la molécule d'ADN, ce qui a pour effet de légèrement dissocier les bases complémentaires de la double hélice à ce niveau. La fixation de TBP est suivie de la fixation des différents TAF qui constituent TFIID. Dans un second temps, le facteur TFIIA (*transcription factor* IIA) se fixe à TBP et stabilise son interaction avec l'ADN. Il semble que sa fixation puisse induire l'expulsion de facteurs inhibiteurs de la transcription. Le troisième acteur est une protéine monomérique appelée TFIIB (*transcription factor* IIB). Le premier tiers de la protéine correspond à un domaine fixant le zinc, dont le rôle est inconnu. Les deux tiers restants interagissent à la fois avec les protéines fixées et à la fois avec l'ADN. Une extrémité se fixe en amont de la boîte TATA (au niveau du grand sillon de l'ADN), l'autre en aval (au niveau du petit sillon de l'ADN). Cette fixation asymétrique pourrait être à l'origine de la détermination du sens de la transcription. Le quatrième acteur est le facteur TFIIF (*transcription factor* IIF) qui est une protéine constituée de quatre sous-unités deux à deux identiques. Il stabilise le complexe créé précédemment, induit des torsions de l'ADN qui facilitent sa fusion au niveau du promoteur. Il possède aussi une grande affinité pour l'ARN polymérase II ce qui permet son association au complexe et vraisemblablement prévient sa fixation à des sites non spécifiques. L'acteur suivant est le facteur TFIIE (*transcription factor* IIE). Les activités de ce facteur sont multiples. Il sépare les deux brins de l'ADN au niveau du promoteur à l'aide d'ATP (adénosine triphosphate). Sa forte affinité pour l'ADN simple brin permet aux deux brins de l'ADN du promoteur de rester bien séparés. Enfin il recrute le sixième et dernier facteur, TFIIH (*transcription factor* IIH), protéine très complexe qui possède 9 sous-unités. Six d'entre elles (XPB, XPD[3], p34, p44, p52 et p62) constituent le « cœur» (*core*) de la protéine. La sous-unité XPD fait le pont entre ce « cœur » et les trois autres sous-unités (Cdk7, cycline H et Mat1) qui constituent le complexe CTD kinase, appelé CAK (*Cyclin-Activating Kinase*), dont l'action est, une fois activée par TFIIE, de phosphoryler le domaine CTD de l'ARN polymérase II (décrit plus haut), étape indispensable à la transition initiation/élongation. La taille du complexe ainsi formé est proche de celle d'un ribosome et il est possible de l'observer au microscope électronique lorsqu'il est attaché à l'ADN.

Remarques :

1 La boîte TATA n'est pas indispensable car la majorité des gènes domestiques (*house keeping*) n'en possèdent pas et des expériences de délétion et de mutation de la boîte TATA n'abolissent pas totalement la transcription (on observe une diminution du taux de transcription et une perte de la fidélité du site d'initiation, c'est-à-dire que la transcription ne démarre plus au niveau d'un point unique, mais à partir de plusieurs points). Le mécanisme est plus complexe au niveau des gènes qui ne possèdent pas naturellement de boîte TATA.

2 TBP semble pouvoir être remplacée par une série de facteurs qui lui ressemblent (*TBP-like*) comme les différents TIC (*TAF and Initiator-dependent Cofactors*). Ces derniers reconnaissent, entre autres, une séquence appelée Inr (*Initiator*) qui est retrouvée dans quelques promoteurs (dont certains ont aussi une boîte TATA).

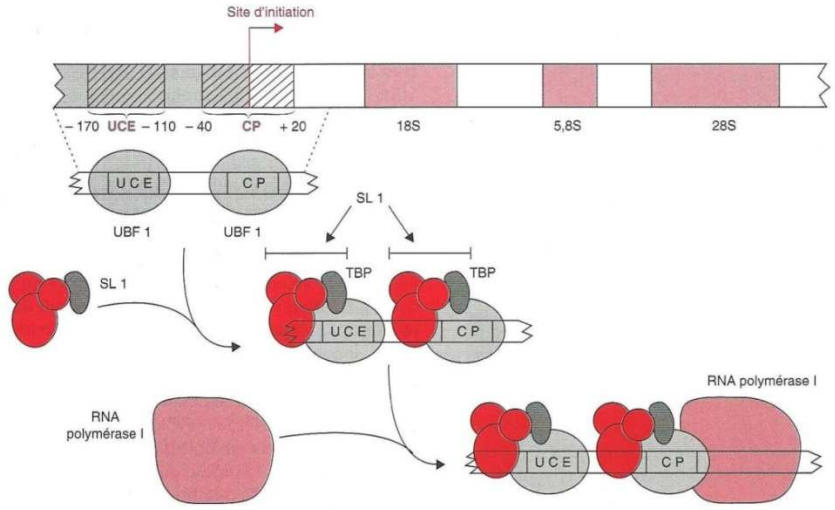
3 Les sous- unités XPB (*Xeroderma Pigmentosum B*) et XPD (*Xeroderma Pigmentosum D*) sont deux ARN-ADN hélicases ATP-dépendantes, indispensables au démarrage de la transcription, qui empêchent les initiations abortives.

Lorsque la polymérase se fixe au complexe d'initiation, le complexe CAK (*Cyclin-Activating Kinase*) phosphoryle le domaine CTD de l'ARN polymérase sur les sérines 2 et 5 de chacune des 52 répétitions qui constituent le domaine. Cette phosphorylation rend cette partie de la polymérase très acide, ce qui induit la polymérase à quitter le site d'initiation et à passer sous la forme d'un complexe d'élongation stable. Des facteurs agissent aussi sur la transition entre initiation et élongation, en permettant notamment à la polymérase de quitter le site d'initiation de la transcription (comme P-TEFb, *Positive Transcription Elongation Factor b*). Dans le même temps de nombreux facteurs qui composaient le complexe d'initiation sont libérés et remplacés par des facteurs d'élongation. Une fois le site d'initiation de transcription libéré, une nouvelle molécule de polymérase peut amorcer un nouveau transcrit. Le nombre de molécules de polymérase engagées dans la transcription d'un gène est proportionnel au taux de transcription de ce gène.

2. L'initiation de la transcription avec l'ARN polymérase I

De la même manière que la Pol II, la Pol I ne se fixe pas directement sur l'ADN, mais par l'intermédiaire de plusieurs facteurs. L'initiation de la transcription avec la Pol I est assez similaire à celle avec la Pol II, et est détaillée en Figure 5.

Figure 5 : Mécanisme d'initiation de la transcription avec l'ARN Pol I



D'après (Kaplan et Delpech, 2007). *In vitro*, la polymérase I n'est pas capable de transcrire les gènes ribosomaux si un extrait nucléaire n'est pas ajouté, ce qui suggère que des facteurs transcriptionnels sont indispensables : il s'agit des facteurs UBF1 (*Upstream-binding factor 1*) et SL1 (*selectivity factor 1*). Ils suffisent à eux seuls pour initier une transcription. Dans un premier temps UBF1 se fixe sur les deux séquences UCE (*Upstream Control Element*) et CPE (*Core Promoter Element*) situées respectivement entre - 170 et - 110 pb en amont du site d'initiation de la transcription d'une part, et entre - 40 et + 20 pb d'autre part. Ces deux séquences particulières possèdent la caractéristique unique pour des séquences promotrices d'être riches en GC. Le facteur SL1 reconnaît alors le complexe et s'y fixe. Ce dernier est un tétramère dont l'une des sous-unités est un polypeptide appelé TBF (TATA *Binding Factor*) qui est trouvé associé à toutes les ARN polymérases eucaryotes (comme décrit plus haut). Comme les promoteurs des gènes codant les ARN des ribosomes n'ont pas de boîte TATA, il est vraisemblable que le rôle de ce facteur est plus de permettre la fixation de l'ARN polymérase I que de reconnaître la séquence du promoteur. Le complexe promoteur/UBF1/SL1 permet à l'ARN polymérase I de se caler strictement au site de début de transcription et d'initier cette dernière. Il est à noter que les gènes qui codent l'ARN des ribosomes sont très nombreux et sont regroupés (cluster). Chaque gène, dont la taille est d'environ 13 kb, est séparé du suivant par un espaceur, d'une taille d'environ 30 kb, constitué de séquences grossièrement répétées. Chez la souris, l'étude de l'ADN situé entre les gènes ribosomaux répétés a montré qu'il contient un autre promoteur suivi d'une séquence stimulatrice (*enhancer*, voir plus loin), constituée d'une suite de séquences répétées, et capable de fixer le facteur de transcription UBF. Le rôle de ces séquences reste à préciser.

3. L'initiation de la transcription avec l'ARN polymérase III

L'une des principales particularités de plusieurs gènes de classe III est que leur promoteur peut être situé dans la partie transcrite. L'ARN polymérase III, qui est une molécule complexe (17 sous-unités chez l'homme pour une masse moléculaire totale de 650 kDa), nécessite également des facteurs protéiques pour agir.

Il existe trois types de promoteurs pour l'ARN polymérase III. L'assemblage de facteurs transcriptionnels est différent pour chacun des types de promoteur. Au cours de leur découverte, les facteurs transcriptionnels correspondants ont été appelés TFIIIA, TFIIIB et TFIIIC. En réalité, les facteurs TFIIIB et TFIIIC sont des complexes protéiques contenant de nombreuses protéines et dont la composition varie suivant le type de promoteur cible (les noms sont identiques mais correspondent à des assemblages de protéines différents). Seul le facteur « de type » TFIIIB est retrouvé pour les trois types de promoteur : il va recruter l'ARN polymérase III dans tous les cas. Le modèle simplifié est le suivant :

- Les promoteurs de type 1 sont associés aux gènes codant les ARN5S du ribosome. Ils sont situés dans la séquence transcrite du gène. Ils fixent le facteur TFIIIA au niveau de la séquence ICR (*Internal Control Région*) située entre les bases + 55 et + 80 par rapport au site d'initiation de la transcription. Le facteur TFIIIA fixé recrute le facteur TFIIIC. Enfin le facteur TFIIIB, qui est constitué de trois protéines (TBP ou TATA *Binding Protein*, Brf1 ou TFIIIB-*related factor 1*, et Bdp1 ou *B double prime 1*), s'associe à ces deux protéines, mais sans interagir avec l'ADN, et recrute l'ARN polymérase III, ce qui initie la transcription.

- Les promoteurs de type 2 sont aussi situés dans la partie transcrite du gène. Ils sont associés aux gènes qui codent les ARNt. Ils fixent le facteur TFIIIC. Cette fixation induit le recrutement du facteur TFIIIB (qui a la même composition que le facteur TFIIIB des promoteurs de type 1), qui recrute l'ARN polymérase III, ce qui initie la transcription.

- Les promoteurs de type 3 ont une organisation très différente de celle de deux autres types. Ils sont associés à plusieurs gènes codant des ARN variés (snARN, certaines RNases,…). Ces promoteurs ressemblent à ceux de l'ARN polymérase II, en partie car ils sont situés en amont de la séquence transcrite. De 3' vers 5' on trouve une boîte TATA, un élément de contrôle proximal (PSE, *Proximal Sequence Element*), et enfin, plus en amont, une séquence de contrôle distale (DES, *Distal Sequence Element*), qui contient une séquence de fixation pour le facteur transcriptionnel Oct1 (*Octamer transcription factor 1*). La distance entre la boîte TATA et la séquence PSE est constante dans tous les gènes. La séquence PSE fait aussi partie du promoteur des autres gènes codant les snARN qui sont transcrits par l'ARN polymérase II. Le système d'initiation au niveau des promoteurs de type 3 est complexe. La séquence DSE fixe le facteur transcriptionnel Oct1 et la séquence PSE fixe le facteur SNAPc (*snRNA activating protein complex*, aussi appelé PTF pour *PSE-binding transcription factor* : ce facteur se fixe aussi sur le promoteur des gènes codant les snARN qui sont transcrits par la polymérase II). Enfin le facteur TFIIIB se fixe grâce à sa sous-unité TBP qui reconnaît la boîte TATA, et interagit aussi avec SNAPc. La composition du facteur TFIIIB utilisé pour les promoteurs de type 3 est légèrement différente de celle des TFIIIB utilisé par les promoteurs de type 1 et 2 (la protéine Brf1 est remplacée par la protéine Brf2). Enfin ce complexe de facteurs transcriptionnels recrute l'ARN polymérase III.

C. Les facteurs de transcription et la régulation transcriptionnelle chez les eucaryotes

D'après (Kaplan et Delpech, 2007).

Chez les eucaryotes, le système de sélection/régulation de l'expression des gènes est complexe et diversifié. En effet, le nombre des nucléotides du génome des cellules eucaryotes est considérablement élevé et l'ADN n'est pas libre mais hautement compacté dans le noyau. Le système de reconnaissance d'une courte séquence par une protéine unique (comme chez les procaryotes) n'est pas envisageable car les constantes d'affinité nécessaires seraient très supérieures à ce que l'on peut rencontrer en biologie et les temps de recherche seraient prohibitifs et sans aucun rapport avec la vitesse et la souplesse de la réponse aux stimuli observée dans les cellules eucaryotes.

De plus chez les organismes eucaryotes complexes comme les mammifères, il existe un fort taux de différenciation cellulaire (les cardiomyocytes, les neurones et les mastocytes par exemple sont des cellules très différentes). Ainsi, toutes les cellules d'un même organisme contiennent en effet la même information génétique (Figure 1), mais tous les gènes ne peuvent pas être transcrits en même temps dans toutes les cellules car la plupart des cellules sont spécialisées dans une fonction. D'où la nécessité chez les eucaryotes d'avoir des mécanismes fins de régulation de l'expression des gènes qui contrôlent l'expression spatiale et temporelle des gènes pour leur permettre de modifier le type et la croissance de leurs cellules dans une grande variété de possibles.

Tout d'abord pour diminuer l'effort de sélection, plusieurs mécanismes épigénétiques mis en place au cours de la différenciation cellulaire (condensation de la chromatine, acétylation des histones, méthylation de l'ADN,…) présélectionnent l'accès de la transcription à certains gènes dans une cellule donnée. En effet, pour pouvoir être transcrit, un gène doit être dans une conformation chromatinienne particulière. À cette présélection s'ajoute une multiplicité de niveaux successifs de régulation (dont quelques-uns sont présentés en Annexe 1) qui permettent un ajustement de la vitesse et de l'intensité de la réaction aux stimuli.

La régulation transcriptionnelle, qui m'intéresse en particulier dans cette thèse, est très complexe et fait intervenir de nombreux protagonistes. Nous venons en effet de voir que la polymérase n'agit pas toute seule mais a besoin de nombreux facteurs, dont des facteurs de transcription. Ces protéines (facteurs *trans*) qui se lient à une séquence particulière (séquence *cis*) participent à l'acquisition (directement ou indirectement) de la spécificité cellulaire évoquée plus haut.

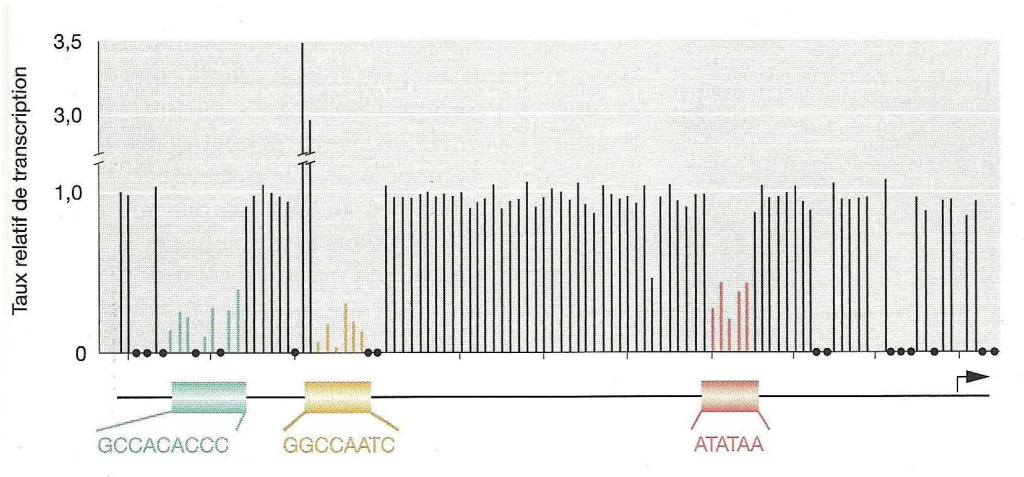
1. Certaines séquences sont capables de modifier le taux de transcription : la régulation en *cis*

Une séquence contiguë à un gène et ayant un effet régulateur sur le taux de transcription de ce gène est appelée élément *cis*-régulateur. Les progrès technologiques (mutagenèse, gène rapporteur,…) ont permis d'étudier avec une très grande précision l'éventuel rôle régulateur de n'importe quelle séquence.

Certaines séquences *cis* ont une localisation parfaitement définie au sein des régions promotrices. *Stricto sensu*, le promoteur correspond à la région où se fixe l'ARN polymérase II, donc à une séquence génomique qui commence un peu avant la boîte TATA et se termine au site d'initiation de la transcription. Mais les séquences nécessaires à la régulation de la transcription chez les eucaryotes remontent beaucoup plus en 5', parfois plusieurs dizaines, voire centaines, de kilobases en amont. La nomenclature utilisée pour désigner ces régions a varié au cours du temps et suivant les auteurs : dans cette thèse, toutes les séquences *cis*-régulatrices situées en 5' du gène

seront appelées séquences régulatrices d'amont. La plupart de ces séquences sont la cible spécifique de facteurs transcriptionnels (ou facteurs *trans*). On y retrouve les boîtes (*box*) CAAT et GC déjà évoquées. Des mutations dans ces sites peuvent avoir un effet important sur la transcription, ce qui démontre leur importance (Figure 6).

Figure 6 : Les éléments proches du promoteur sont nécessaires à une transcription efficace

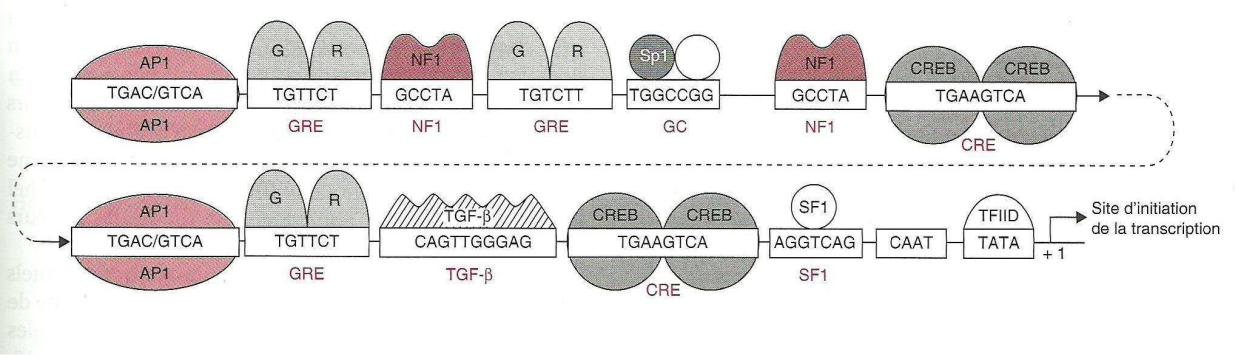


D'après (Griffiths *et al.*, 2013). Il s'agit de l'expérience presque historique démontrant l'importance de certaines séquences régulatrices d'amont : des auteurs (Maniatis *et al.*, 1987) ont déterminé l'effet sur le niveau de transcription de mutations ponctuelles dans le promoteur et les éléments proches du promoteur du gène de la β-globine. Chaque trait représente le taux de transcription relatif (par rapport à un promoteur de type sauvage) lorsque l'on mute une des 109 bases testées en amont du premier exon du gène. Les substitutions de bases modifiant significativement le taux de transcription se trouvent à l'intérieur de trois éléments : la boîte TATA (en rouge, à environ - 25 pb), la boîte CAAT (en jaune, à environ - 70 pb) et la boîte GC (en vert, à environ - 70 pb, mais peut parfois se trouver entre les deux premières boîtes). Les positions signalées par un point noir n'ont pas été testées.

Une caractéristique de ces séquences est qu'elles sont le plus souvent grossièrement symétriques, les protéines qui s'y fixent étant en général des homo- ou des hétéro-dimères. Le fait qu'un gène puisse répondre ou pas à un stimulus donné ou présenter une spécificité tissulaire d'expression dépend en partie de la présence ou de l'absence de ces séquences (Figure 7). Ainsi la présence d'un ERE (*Estrogen Responsive Element* ou *Estrogen Response Element*) en amont d'un gène permet à celui-ci d'avoir une expression modulée par les œstrogènes, mais seulement dans les cellules qui possèdent le récepteur des œstrogènes (qui est un facteur *trans*), et lorsque l'hormone est présente. Dans les cellules ne possédant pas le récepteur ou en l'absence de l'hormone, la séquence est présente mais sans effet.

Certains gènes, comme celui de l'alpha-amylase, possèdent plusieurs promoteurs possibles de force inégale. Il en résulte que le type de messager transcrit et le taux de sa transcription vont dépendre du promoteur utilisé. Le choix ne se fait pas au hasard, mais résulte de l'action des facteurs de transcription dont certains sont spécifiques de tissu, ce qui explique pourquoi certains tissus expriment un type de messager et d'autres tissus un autre type. Cette régulation s'appelle la régulation par choix de promoteur (Figure 8).

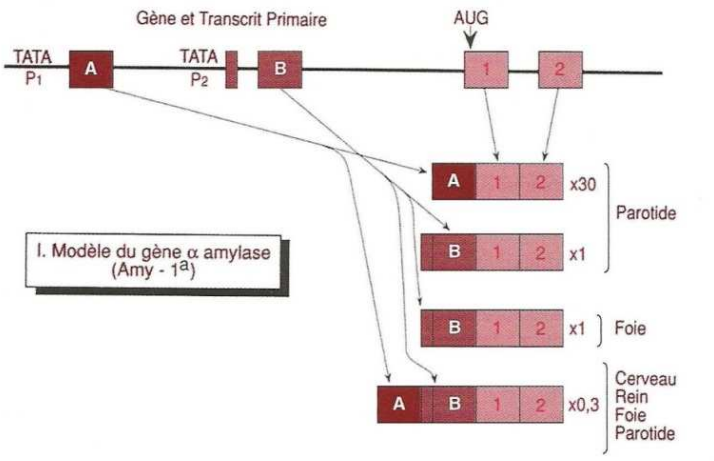
Figure 7 : Organisation de la région promotrice du gène de l'aromatase



D'après (Kaplan et Delpech, 2007). La région promotrice du gène de l'aromatase possède plusieurs séquences *cis*-régulatrices qui confèrent au gène un panel tissulaire d'expression particulier.

AP1 : *Activator Protein 1*, facteur de transcription formé par l'hétérodimérisation de protéines des familles c-Jun et c-Fos impliquées dans la régulation du cycle cellulaire (oncoprotéines).
GRE : *Glucocorticoide Responsive Element*
G : récepteur nucléaire aux glucocorticoïdes
NF1 : *Nuclear Factor 1*, facteur se liant à la boîte CAAT.
Sp1 : *Specificity protein 1*, protéine qui reconnaît spécifiquement la séquence GGGCGG (parfois appelée GC *box*), séquence promotrice de nombreux gènes de ménage.
CREB : *Cyclic-AMP Response Element-binding protein*.
CRE : *Cyclic-AMP Response Element* (élément de réponse à l'AMPc).
TGF-β : *Transforming Growth Factor beta*. Le schéma n'indique pas que le TGF-β se fixe sur l'ADN, mais seulement qu'il existe un élément de réponse (pouvant être appelé abusivement un « élément de réponse au TGF- β ») sur lequel se fixe un facteur activé via la voie du TGF- β.
SF1 : *Steroidogenic Factor 1*.
TFIID : *Transcription Factor IID* (facteur général de la transcription).

Figure 8 : Diversification des transcrits par utilisation de promoteurs alternatifs



D'après (Kaplan et Delpech, 2007). Le gène de l'alpha-amylase possède deux promoteurs, P1 et P2. Leur utilisation alternative ou simultanée combinée à un épissage alternatif est spécifique de tissu et conduit à un taux d'expression variable suivant l'organe.

Il existe des banques de données spécifiquement dédiées à ces séquences, comme EPD (*Eukarytic promotor Database :* www.epd.isb-sib.ch). Des programmes informatiques simples permettent de les rechercher dans n'importe quelle séquence d'ADN. Cette recherche est particulièrement utile lorsque l'on vient d'isoler un gène et que l'on s'intéresse à la régulation de son expression. Comme maintenant la totalité de la séquence du génome humain est connue, les analyses bioinformatiques permettent aussi de repérer l'ensemble des gènes qui possèdent des séquences régulatrices homologues et qui pourraient donc être soumis au même mécanisme de contrôle de leur expression. Il convient cependant d'être très prudent car il existe dans le génome de nombreuses séquences qui ressemblent à des séquences cibles de facteurs de transcription mais qui ne sont associés à aucun gène et qui ne sont pas fonctionnelles. Le caractère régulateur d'une séquence ne peut être affirmé que lorsqu'il a été confirmé expérimentalement.

D'autres séquences *cis* pouvant avoir des localisations variées ont été regroupées sous le terme de séquences stimulatrices de transcription (*enhancers*). Ces séquences possèdent en commun une série de propriétés qui les définissent :

- la principale est d'augmenter considérablement le taux de transcription du gène auquel elles sont associées ;

- elles peuvent êtres localisées en amont (en 5') ou en aval (en 3') du gène, ou même dans un intron du gène (Levine et Tjian, 2003) ;

- l'éloignement par rapport au gène contrôlé peut être de quelques centaines, voire milliers de paires de bases (Levine et Tjian, 2003) ;

- en général leur inversion ne se traduit pas par la perte de leur effet sur la transcription, mais simplement par une légère diminution de cet effet ;

- elles gardent leur caractère activateur lorsqu'on les déplace (en 5', en 3', ou même dans le gène), mais l'effet activateur est maximal en un point donné, et devient d'autant plus faible que l'on s'éloigne de cet emplacement normal.

Un *enhancer* typique peut faire jusqu'à 500 paires de bases de long et contenir plusieurs sites de liaison pour au moins deux ou trois facteurs de transcription différents (Levine et Tjian, 2003). La première séquence stimulatrice de transcription de cellule eucaryote mise en évidence a été celle des immunoglobulines. Leur mécanisme d'action commence à être déchiffré. La plupart agissent par l'intermédiaire de protéines (facteurs *trans*, voir plus loin), notamment celles qui n'ont d'effet que dans certains organes. Il s'agit très souvent d'*enhancers* spécifiques d'un tissu ou d'un type cellulaire (Levine et Tjian, 2003). Certaines agissent aussi, semble-t-il, sans l'intermédiaire de protéines, en modifiant la structure spatiale de l'ADN ou son taux de torsion. Enfin leur activité peut aussi impliquer des modifications épigénétiques, comme des méthylations de l'ADN. Des séquences du même type, agissant suivant les mêmes mécanismes mais ayant un effet inverse sur la transcription, ont été caractérisées : elles sont appelées séquences extinctrices (*silencers*).

La liste des séquences cis-régulatrices s'allonge d'année en année. En effet, au début des recherches sur le génome humain, les scientifiques se sont rendu compte que l'information contenue dans l'ADN était noyée dans une quantité considérable de séquences non-codantes (seulement 2 % du génome environ code pour des protéines) qu'ils ont appelé « ADN poubelle » (car elles semblaient inutiles). Depuis, la comparaison entre une douzaine d'espèces de longues séquences (350 kb) situées en amont de quelques gènes a mis en évidence l'existence de nombreuses courtes séquences (une centaine de paires de bases) encore plus conservées que les séquences exoniques. Une telle conservation est en général corrélée à un rôle important et le rôle régulateur de certaines

d'entre elles a pu être montré en les insérant dans des vecteurs d'expression contenant un gène rapporteur où elles entraînent une activation de la transcription. Mais toutes n'ont pas d'effet sur la transcription : le rôle de ces séquences très conservées reste donc à être largement précisé. Pour aller plus loin dans cette voie, l'Institut national de recherche sur le génome humain (NHGRI, *National Human Genome Research Institute*) a lancé en Septembre 2003 un grand projet de recherche publique nommée ENCODE (*Encyclopedia Of DNA Elements*, http://www.encodeproject.org/ENCODE/), visant à lister et annoter tous les éléments fonctionnels dans le génome humain (c'est-à-dire toute séquence codant pour une protéine ou un ARN non-codant ou qui a une propriété particulière, par exemple de se lier à un facteur transcriptionnel). Ce grand projet a pour but d'améliorer les connaissances sur la chromatine, la régulation des gènes, mais aussi les éléments en relation avec les maladies (certaines maladies sont dues à des mutations dans la région promotrice, dans un site de fixation d'un facteur de transcription ou dans un intron, c'est-à-dire dans une région non-codante). Les premiers résultats de ce projet ont déjà révélé que 80 % du génome aurait une fonction biochimique (ENCODE Project Consortium *et al.*, 2012) et que 75 % du génome serait transcrit (Djebali *et al.*, 2012). Certains journaux titrent déjà : « À quand la fin de l'ADN poubelle ? » (Rosier, 2012).

En résumé, les séquences régulatrices d'amont confèrent à un gène des potentialités de régulation. Cette régulation (qui peut être positive ou négative) est assurée par les protéines qui s'y fixent, et qui sont des facteurs *trans*.

2. Certaines protéines sont capables de modifier le taux de transcription : la régulation en *trans*

La plupart des séquences évoquées au paragraphe précédent ne modifient pas seules le taux de transcription. Des protéines interagissant avec elles, appelées facteurs de transcription, sont responsables de la modification observée. Ce type de régulation porte le nom de régulation en *trans*.

L'un des premiers facteurs de transcription à avoir été caractérisé, purifié et cloné est la protéine Sp1, qui reconnaît la séquence GGGCGG (parfois appelée *GC box*). Depuis Sp1, la liste des facteurs de transcription caractérisés s'allonge d'année en année.

Au début de leur caractérisation, on a mis en évidence deux grands types de facteurs transcriptionnels : ceux qui se fixent à l'ADN en y reconnaissant une séquence cible et ceux qui assurent des interactions entre les acteurs de la transcription, notamment entre les facteurs qui se fixent à l'ADN et les complexes généraux de transcription associés à l'ARN polymérase II (décrits précédemment). Pour des raisons pratiques, les premiers ont gardé le nom de facteur de transcription alors que les seconds ont été appelés co-facteurs. Les facteurs de transcription sont mieux connus car il est plus facile de caractériser techniquement des interactions ADN/protéine que des interactions protéines/protéines spécifiques dans des complexes qui impliquent des dizaines de protéines.

Une définition commune d'un facteur de transcription est une protéine contenant un domaine spécifique de liaison à l'ADN (*DNA-binding domain*) et régulant la transcription d'un (ou plusieurs) gène(s)-cible(s). Les données accumulées ont permis de mettre en évidence quelques caractéristiques partagées par la plupart des facteurs de transcription :

- Le domaine de liaison à l'ADN est spécifique et possède une structure bien particulière. C'est souvent à partir de la structure de ce domaine que les facteurs de transcriptions ont été classés en familles (voir plus loin).

- Le motif de reconnaissance sur l'ADN est d'environ 6 à 10 paires de bases seulement. Le facteur de transcription ne reconnaît pas forcément une séquence particulière mais peut reconnaître un ensemble de séquences regroupées sous le terme de séquence consensus.

- Il s'agit le plus souvent d'hétérodimères.

- Ils peuvent influer positivement ou négativement sur la transcription du gène-cible, en fonction de la présence d'autres domaines fonctionnels sur la protéine et de l'impact global de l'ensemble du complexe (Phillips et Hoopes, 2008).

- Chaque monomère de ces hétérodimères possède au minimum deux domaines : le domaine d'interaction avec l'ADN et le domaine d'activation de la transcription. Les expériences de construction-transfection ont montré que les deux types de domaines sont interchangeables entre les différents facteurs de transcription. Par exemple, une protéine constituée d'un domaine d'activation de transcription provenant d'un facteur transcriptionnel de mammifère couplé à un domaine de fixation à l'ADN provenant d'une levure, est parfaitement fonctionnelle et activera tous les gènes possédant la séquence cible du domaine de fixation à l'ADN.

- Ils peuvent être plus complexes et posséder un ou plusieurs domaines fonctionnels supplémentaires facultatifs, comme des domaines d'interaction avec d'autres facteurs de transcription, avec des protéines co-activatrices (ou co-facteurs), avec l'ARN polymérase II, avec des complexes de remodelage de la chromatine, avec de petits ARN non codants, ou des domaines de modulation d'activité sous l'effet d'un effecteur (par exemple une hormone). Des molécules de signalisation (AMPc ou hormones par exemple) peuvent influencer l'activation de facteurs de transcription en se liant de façon covalente ou en modifiant leurs domaines fonctionnels. Contrairement aux précédents, ces domaines ne présentent pas de structures typiques.

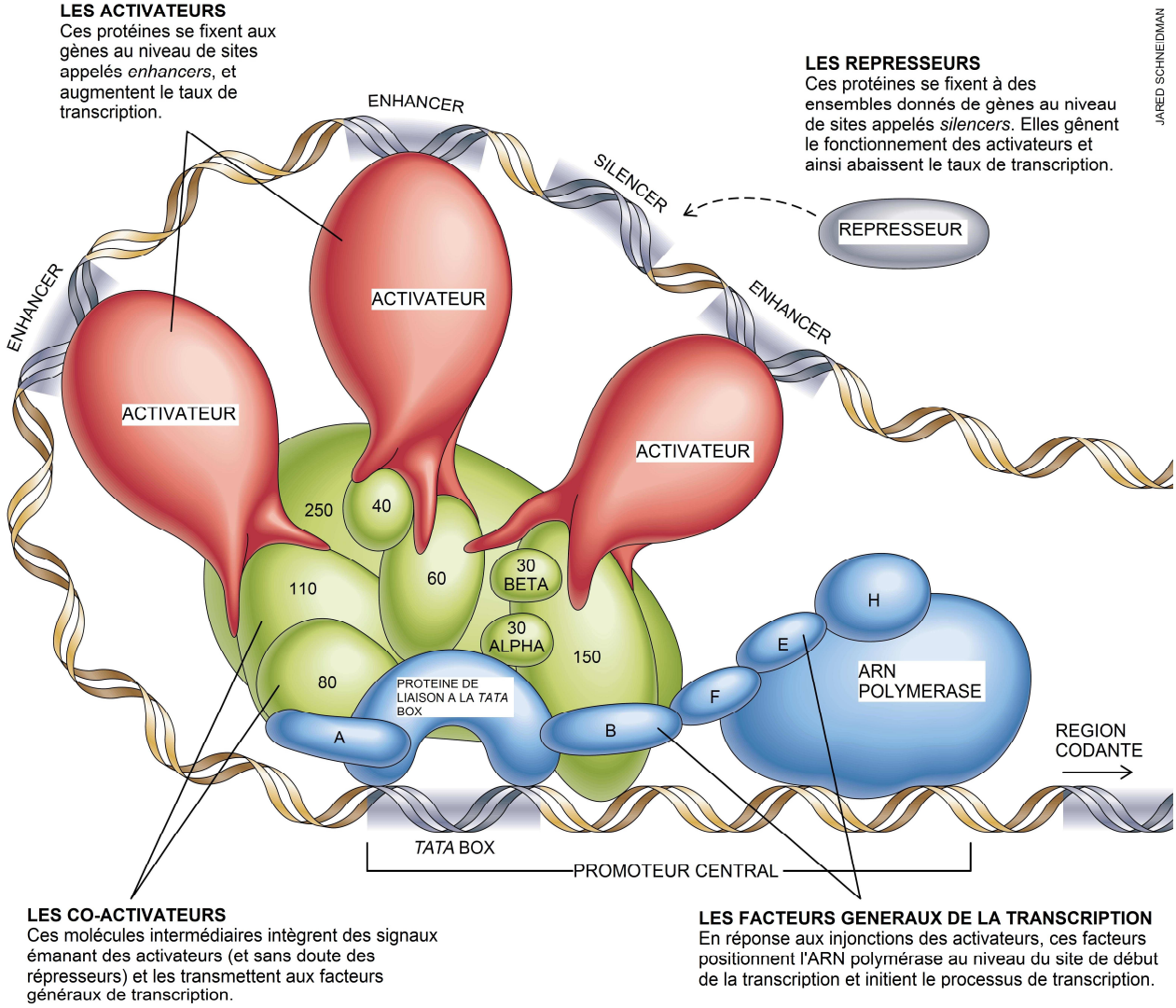
- Ils ne sont pas, en général, spécifiques d'un seul gène-cible (contrairement aux procaryotes). En effet chez les eucaryotes, chaque gène possède son propre système de régulation qui peut mettre en jeu jusqu'à plusieurs dizaines de facteurs. Si donc chaque facteur transcriptionnel était spécifique d'un seul gène, comme l'expression de chacun de ces facteurs devrait être elle aussi régulée, d'autres facteurs seraient nécessaires, et ainsi de suite, et le nombre de facteurs croîtrait de manière exponentielle (ce qui est impossible). La diversité des réponses observées ne résulte donc pas de la multiplicité des facteurs de transcription mais de la combinatoire d'un nombre limité de facteurs de transcription et de séquences cis (le principe est semblable à celui du langage, de l'écriture ou de la musique où un nombre limité de sons ou de symboles permet une infinité de combinaisons différentes). Plusieurs facteurs de transcription sont connus pour faciliter la transcription sur des centaines de promoteurs différents, alors que certains ne sont actifs que sur quelques-uns seulement.

3. Mécanismes de la régulation transcriptionnelle par les facteurs de transcription

Les connaissances sur les mécanismes de la régulation de la transcription par les facteurs de transcription, leur structure et leur mode d'action, ont considérablement progressé au cours des dernières décennies. Les régions régulatrices d'amont et les différents facteurs s'y fixant commencent à être connus pour de nombreux gènes. Il est clair maintenant que les régions régulatrices en amont des gènes sont constituées d'une série de modules, chaque module conférant au gène une possibilité supplémentaire de régulation (Figure 7). Mais il est certain aussi que la simple interaction d'une série de facteurs avec des séquences d'ADN n'est pas suffisante. Plusieurs résultats (notamment des expériences de mutagenèse dirigée) montrent que la disposition relative dans l'espace des différentes protéines est plus importante que leur distance par rapport au

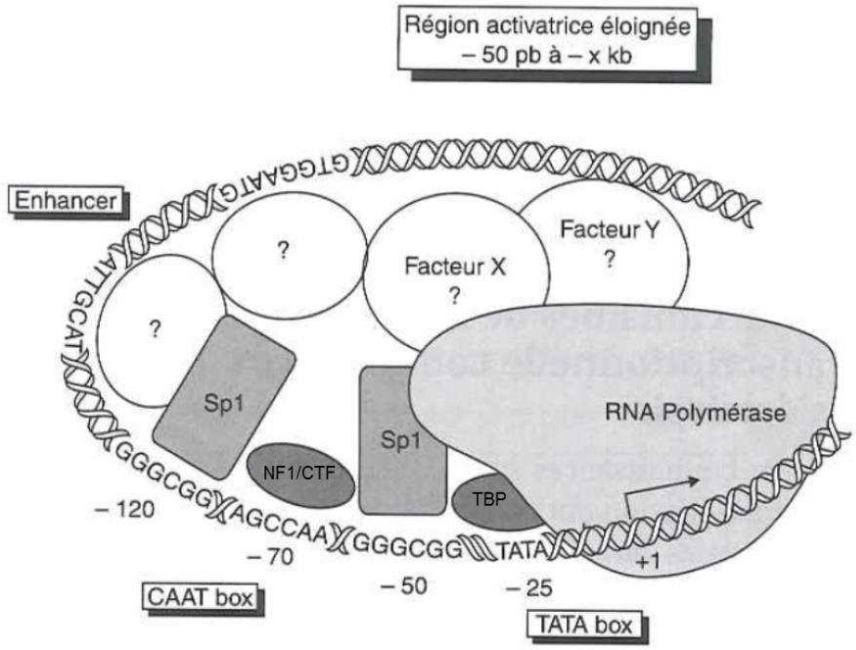
promoteur et sont fortement en faveur du rôle de la structure tridimensionnelle de la chromatine et des interactions protéine-protéine dans la régulation : plusieurs schémas de cette idée existent, le premier date de 1995 (il a d'ailleurs peu évolué depuis) et est présenté en Figure 9. De même, il a été démontré que des séquences situées très en amont (plusieurs dizaines de kilobases) sont nécessaires à la bonne régulation de certains gènes. L'une des hypothèses les plus vraisemblables est que les repliements de la chromatine rapprochent ces séquences du point d'initiation de la transcription. Ainsi chaque gène est précédé par une série de séquences régulatrices d'amont assemblées comme un jeu de Lego® et séparées par des séquences non critiques dont le seul but est, semble-t-il, de placer stériquement la chromatine dans des conditions optimales pour que les protéines qui s'y fixent aient une configuration particulière dans l'espace, puissent interagir (même quand ils sont liés à des séquences d'ADN de plusieurs centaines de paires de bases d'intervalle) et jouent leur rôle dans la transcription (Figure 10).

Figure 9 : Schéma de la régulation de la transcription par les facteurs de transcription



D'après (Tjian, 1995). L'assemblage moléculaire contrôlant la transcription dans les cellules eucaryotes est constitué de quatre types de composants (les protéines numérotées sont les noms des sous-unités de l'ARN polymérase II, chaque sous-unité est dénommée d'après sa masse moléculaire en kDa). Les facteurs généraux de transcription (notés A, B, F, E et H) sont essentiels à la transcription mais ne peuvent par eux-mêmes augmenter ni abaisser le taux de la transcription. Cette tâche revient aux molécules régulatrices appelées activateurs ou répresseurs. Les activateurs, et sans doute les répresseurs, communiquent avec les facteurs généraux par le biais de co-activateurs (des protéines complexées avec la protéine qui se fixe à la boîte TATA, le premier des facteurs généraux de transcription à se fixer au promoteur central).

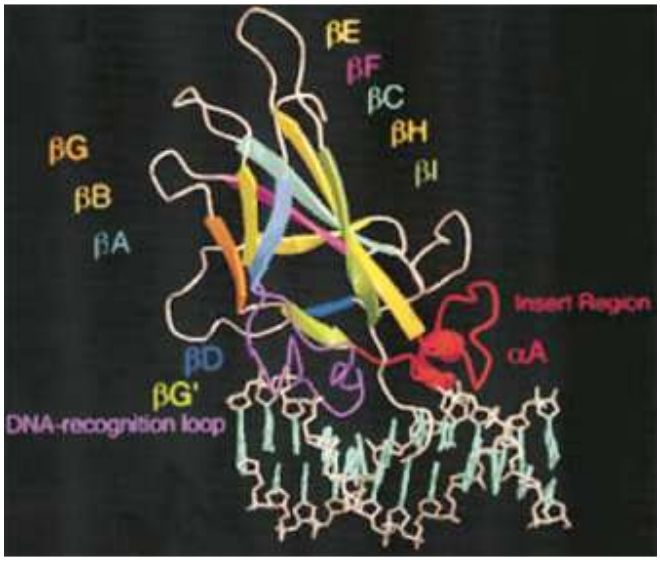
Figure 10 : Représentation schématique du mode d'action des facteurs transcriptionnels sur la régulation de l'initiation de la transcription



D'après (Kaplan et Delpech, 2007). De manière schématique, chez les eucaryotes, l'essentiel de la régulation de la transcription résulte de l'interaction de séquences localisées en 5' du point d'initiation, parfois très éloignées (éléments *enhancers* par exemple), et d'un complexe protéique tel celui schématisé dans la figure. Une première série de protéines, dont la plupart sont plus ou moins ubiquitaires, se fixent au niveau de séquences spécifiques dans et en amont du promoteur ; les repliements de la chromatine, l'interaction entre ces protéines et éventuellement avec des protéines non fixées à l'ADN créent une structure qui permet à l'ARN polymérase de se fixer et ainsi d'initier la transcription. De la facilité avec laquelle cette initiation pourra se faire et de la vitesse de la transcription dépendra le taux apparent de transcription.
Sp1 (*Specificity protein 1*) est une des premières protéines *trans*-régulatrice à avoir été caractérisée : c'est une protéine en doigt de zinc qui reconnaît spécifiquement la séquence GGGCGG (parfois appelée GC *box*), séquence promotrice de nombreux gènes de ménage.
TBP (TATA *Binding Protein*) : facteur se liant à la boîte TATA.
NF1 (*Nuclear Factor 1*) ou CTF (CAAT *box transcription factor*) : facteur se liant à la boîte CAAT.
« Facteur X », « Facteur Y », « ? » = facteurs hypothétiques.

Ainsi, le mode d'action d'un facteur de transcription est le suivant : d'abord, il reconnaît et se lie à un segment d'ADN dans le promoteur et/ou la région *enhancer*. Souvent la liaison à l'ADN s'accompagne d'une modification de la conformation ou de la structure tridimensionnelle du facteur de transcription. Par exemple, les deux boucles dans NFATc1 qui interagissent avec l'ADN se trouvent dans des conformations différentes, selon que NFATc1 est complexée avec l'ADN ou non (Figure 11). Puis, il interagit généralement avec d'autres facteurs de transcription : par exemple, deux facteurs de transcription liés à des sites près l'un de l'autre sur le brin d'ADN peuvent se combiner pour former un dimère et plier l'ADN d'une certaine façon. Plusieurs facteurs de transcription peuvent s'accumuler, créant un complexe de la taille d'un ribosome. Une fois tous ces facteurs liés ensemble, les modifications liées aux domaines fonctionnels du facteur de transcription et/ou les interactions covalentes avec d'autres facteurs peuvent activer ou réprimer la transcription, selon qu'ils permettent ou bloquent le recrutement de l'ARN polymérase. Ainsi, le contrôle de la transcription dépend de l'interaction de l'ensemble des facteurs de transcription et de l'action qu'ils ont ensemble sur l'ARN polymérase : recrutement ou blocage (Phillips et Hoopes, 2008).

Figure 11 : Structure tridimensionnelle du complexe NFATC1-ADN



D'après (Zhou *et al.*, 1998). Représentation topologique d'éléments de structure secondaire du complexe formé par le facteur de transcription NFATc1 et sa séquence de liaison de 12 paires de bases à l'ADN. Le complexe NFATc1-ADN montre que NFATc1 est une structure en feuillets bêta antiparallèles composée de dix brins. Les changements les plus radicaux qui se produisent lors de la liaison à l'ADN impliquent deux grandes boucles de surface (légende d'après (Phillips et Hoopes, 2008).

4. Place des facteurs de transcription dans la régulation transcriptionnelle

Le mécanisme est en réalité plus complexe lorsqu'on ajoute les autres acteurs de la régulation transcriptionnelle (co-activateurs, complexes de remodelage de la chromatine, enzymes de modification des histones…). La fixation de multiples protéines régulatrices sur les nombreux sites de liaison présents dans un enhancer, associées à ses nombreuses autres protéines, peut catalyser la formation d'un enhanceosome (Griffiths *et al.*, 2013), un gros complexe protéique qui active la transcription de manière synergique grâce aux différentes actions des divers acteurs (le mécanisme complexe de ce complexe ne sera pas détaillé dans cette thèse).

Les mécanismes impliqués dans la régulation de la transcription sont à la base de la décision de transcrire un gène dans une cellule donnée. On estime que 5 à 10 % de la capacité codante des métazoaires est dédiée aux protéines qui régulent la transcription (Levine et Tjian, 2003) : parmi elles, on retrouve les facteurs généraux de la transcription (en comptant l'ARN polymérase), les complexes modifiant et remodelant la chromatine (acétylation des histones, méthylation des histones, méthylation de l'ADN,… que je ne développerai pas dans cette thèse) et les facteurs de transcription possédant un domaine spécifique de liaison à l'ADN. L'expression spécifique de nos gènes dans les cellules est donc régulée en grande partie par des facteurs de transcription, qui sont à la base de la différenciation cellulaire (Levine et Tjian, 2003), mais la complexité et les nuances fines de l'expression de l'ADN chez les eucaryotes résulte d'une lecture combinée de l'état de la chromatine et des signaux de plusieurs facteurs de transcription interagissant ensemble, plutôt que d'une lecture des signaux des facteur de transcription seuls (Phillips et Hoopes, 2008).

De plus, l'activité de certains facteurs de transcription peut être modulée. En effet, Dans certaines circonstances, la cellule doit pouvoir répondre de manière immédiate à un stimulus. Si cette réponse met en jeu l'activation de gènes, le facteur transcriptionnel nécessaire doit être disponible immédiatement. Pour cela il est indispensable qu'il soit déjà présent dans la cellule mais sous une forme inactive. De nombreux processus peuvent être responsables de cette activation, notamment des stimuli extra- ou intracellulaires (hormones, AMP cyclique, ions, choc thermique,

certains lipides...). Les séquences *cis* de ce type sont appelées RE (*Responsive Element*), une lettre supplémentaire étant pour indiquer l'effecteur (qui peut aussi bien être activateur qu'inhibiteur) : par exemple GRE (*Glucocorticoid Responsive Element*) pour l'élément de réponse aux glucocorticoïdes, ERE (*Estrogen Responsive Element*) pour la réponse aux œstrogènes, ou CRE (*Cyclic AMP Responsive Element*) pour la réponse à l'AMPc.

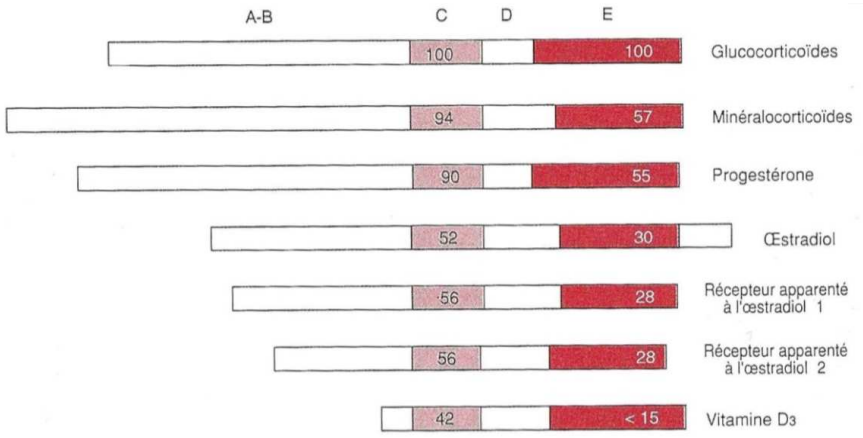
- La phosphorylation peut modifier l'activité d'un facteur transcriptionnel : un exemple de ce mécanisme se trouve dans les facteurs transcriptionnels de la famille CREB (*Cyclic AMP Responsive Element Binding protein*), qui se fixent sur le motif palindromique CRE (TGACGTCA). L'AMPc est un second messager hormonal dont on sait depuis plusieurs années qu'il exerce une grande part de son rôle régulateur sur les métabolismes en activant des protéines kinases regroupées sous le nom de famille Protéine Kinase A (PKA). Mais il a aussi été démontré que l'AMPc jouait un rôle dans la régulation de certains gènes impliqués dans des voies métaboliques aussi différentes que le métabolisme des glucides, la neurotransmission, la réponse immunitaire, la croissance cellulaire ou le rythme circadien. Les gènes dont l'expression est sous le contrôle de l'AMPc possèdent dans leur région régulatrice d'amont une ou plusieurs séquences CRE fonctionnelles. Le facteur CREB est trouvé sous forme de monomère (forme inactive) et de dimère (forme active qui peut se fixer à une séquence CRE). Des expériences de cotransfection d'un gène rapporteur associé à la séquence CRE ont montré que l'activation de l'expression du gène rapporteur nécessitait une phosphorylation de la protéine CREB sur la sérine 133 par la protéine kinase A (résultat qui a été confirmé par mutagenèse). Cette sérine particulière se trouve sur le domaine KID (*Kinase Inductible Domain*) de CREB, qui possède un site de fixation pour une protéine appelée CBP/p300 (*CREB Binding Protein*, il s'agit en réalité de deux protéines différentes mais leurs rôles sont si similaires qu'il est impossible de les distinguer sans utiliser des modèles transgéniques où leurs gènes respectifs ont été invalidés). Le rôle de CBP/p300 (confirmé par les expériences de transgénèse) est fondamental dans la transcription : il est de faire le pont entre un facteur fixé sur l'ADN à distance du promoteur, et le complexe d'initiation. CBP/p300 possède des centaines d'autres partenaires que CREB, comme des facteurs généraux de la transcription (TFIID, TFIIB,…), l'ARN polymérase II et des médiateurs (*mediators*). De plus CBP/p300 possède une activité histone acétyltransférase (HAT) qui permet de faciliter l'accès des facteurs transcriptionnels à la région promotrice. Pour que CREB agissent et interagisse avec CBP, il faut donc que KID soit phosphorylé par la kinase A (activée au préalable par l'AMPc) sur sa sérine 133. Les phosphorylations étant des processus extrêmement rapides, la réponse transcriptionnelle est également rapide. Même si la stimulation par l'AMPc persiste, une baisse de l'activité de CREB a lieu au bout d'environ une heure, et est due à la déphosphorylation de la sérine 133 de CREB par les phosphatases (Kaplan et Delpech, 2007).

- La liaison à un ligand peut modifier l'activité d'un facteur transcriptionnel : l'association d'un facteur de transcription à un ligand peut être à l'origine :

- Soit de sa séquestration dans le cytoplasme. Dans ce cas, bien que présent, le facteur ne pourra exercer son effet puisqu'il ne peut atteindre sa cible sur l'ADN qui est nucléaire. La dissociation du complexe à la suite d'un stimulus permet au facteur de migrer dans le noyau pour y jouer son rôle, et ceci en un temps très court. Il s'agit par exemple du facteur NFκB (*nuclear factor kappa-light-chain-enhancer of activated B cells*, acteur majeur dans la transmission des signaux impliqués dans l'inflammation) qui, dans le cytoplasme, fixe une autre protéine, IKB (*Inhibitor of κB*), ce qui a pour effet d'empêcher sa translocation dans le noyau (forme inactive). La phosphorylation d'IKB à la suite d'un stimulus (antigène, interleukines,...) provoque la dissociation du complexe et permet aux molécules de NFκB libérées de s'associer deux à deux pour constituer des tétramères (forme active) qui migrent vers le noyau pour se fixer sur leur séquence cible (GGGGACTTTCC) et stimuler la transcription de gènes (comme les gènes des chaînes k des immunoglobulines notamment) dans leur région régulatrice d'amont.

- Soit de son activation et de sa translocation dans le noyau. Il s'agit par exemple de la superfamille des récepteurs nucléaires d'hormones. Depuis longtemps on savait que les hormones stéroïdes et thyroïdiennes agissent dans la cellule par l'intermédiaire d'un récepteur intracellulaire, et entraînent la synthèse de messagers spécifiques. Tous ces récepteurs possèdent une structure globale identique schématisée dans la Figure 12. Plus de 60 gènes codant ces récepteurs sont maintenant caractérisés et de nouveaux récepteurs sont progressivement ajoutés à la liste d'année en année, notamment des récepteurs orphelins, c'est-à-dire dont on ne connaît pas le ligand. Parmi les récepteurs orphelins, certains pourraient ne pas avoir de ligand et donc posséder une action constitutive. La dissection en domaines distincts de ces récepteurs a pu être effectuée grâce à des expériences élégantes de transfection et gène rapporteur (voir plus loin), ayant permis d'affirmer la réalité biologique du modèle. Les mécanismes d'action impliqués sont différents suivant les types de récepteurs. Certains fonctionnent sous forme d'homodimères d'autres sous forme d'hétérodimères. Tous sont strictement nucléaires, à l'exception du récepteur aux glucocorticoïdes. En l'absence d'hormone, leur fixation sur l'ADN au niveau de leurs séquences cibles se traduit par un recrutement d'histone désacétylases (HDAC) : la désacétylation des histones des nucléosomes de la chromatine environnante entraîne une inhibition de la transcription. La fixation de l'hormone sur le récepteur induit une transconformation allostérique du récepteur qui, dans sa nouvelle forme, libère les désacétylases, recrute des histones acétyltransférases (HAT) et démasque un site de fixation de facteurs transcriptionnels activateurs, les co-activateurs, au niveau du domaine C-terminal (domaine E sur la Figure 12). La fixation de ces derniers pourrait agir en stabilisant le récepteur dans sa forme active. Si les autres partenaires nécessaires sont présents, il en résulte une forte activation de la transcription et la réponse est très rapide. Pour le cas du récepteur aux glucocorticoïdes, en l'absence de l'hormone, il reste séquestré dans le cytoplasme du fait de son association, en très grands complexes, à la protéine de choc thermique hsp90 (*heat shock protein 90*). La fixation de l'hormone entraîne la dissociation du complexe et la migration du récepteur vers le noyau qui se fixe sur un motif GRE (*Glucocorticoide Responsive Element*). Le mécanisme d'activation de la transcription est similaire aux cas précédents.

Figure 12 : Structure des récepteurs nucléaires d'hormones



D'après (Kaplan et Delpech, 2007). La structure et les homologies de séquence protéique des différents récepteurs sont déduites de la séquence de leurs gènes. Les chiffres indiquent le pourcentage d'homologie de séquence des acides aminés en prenant arbitrairement le récepteur des glucocorticoïdes comme référence. La structure globale de tous ces récepteurs est la suivante :

- Une extrémité NH_2 terminale de longueur variable, non conservée et spécifique du récepteur. On ne connaît pas le rôle fonctionnel de ce domaine (domaine A-B du gène, en blanc).
- Une région d'environ 65 acides aminés, très conservée (90 % d'identité de séquence entre le récepteur des glucocorticoïdes et celui de la progestérone, 42 % d'identité entre les récepteurs les plus éloignés). Ce domaine conservé (domaine C du gène, en rose) est celui qui interagit avec l'ADN. On y retrouve deux motifs en doigt de gant

(Cys_4, voir plus loin), le premier étant le plus conservé. Il est à noter que la région qui relie ces deux doigts de gant est elle aussi très conservée et joue un rôle majeur dans la reconnaissance de la séquence cible sur l'ADN.

- Une région non conservée de longueur variable (domaine D, en blanc) qui joue un rôle de charnière entre les domaines C et E, permettant une transconformation du récepteur lorsque l'hormone se fixe, et qui contient des signaux de localisation du récepteur dans le noyau.
- Le récepteur se termine par un domaine de longueur variable qui correspond à la zone où se fixe l'hormone (domaine E, en rouge). Bien que les hormones puissent être de structures extrêmement différentes, cette région est relativement bien conservée (entre 15 et 57 % d'identité de séquence en acides aminés). Ce domaine est aussi celui qui module la transcription et qui fixe les co-activateurs.

Il est à noter enfin, que le contrôle de l'initiation de la transcription n'est pas absolu : en effet, le développement de la technique PCR a permis de montrer que l'efficacité des mécanismes de régulation de la transcription n'est pas absolue et que tous les gènes sont en fait transcrits à taux ultra-faible dans toutes les cellules. Ce phénomène, dont on ne connaît pas le mécanisme, a été appelé transcription illégitime. Il est particulièrement utile pour le biologiste car il lui permet d'étudier n'importe quel transcrit dans n'importe quelle cellule, ce qui n'était pas évident au départ.

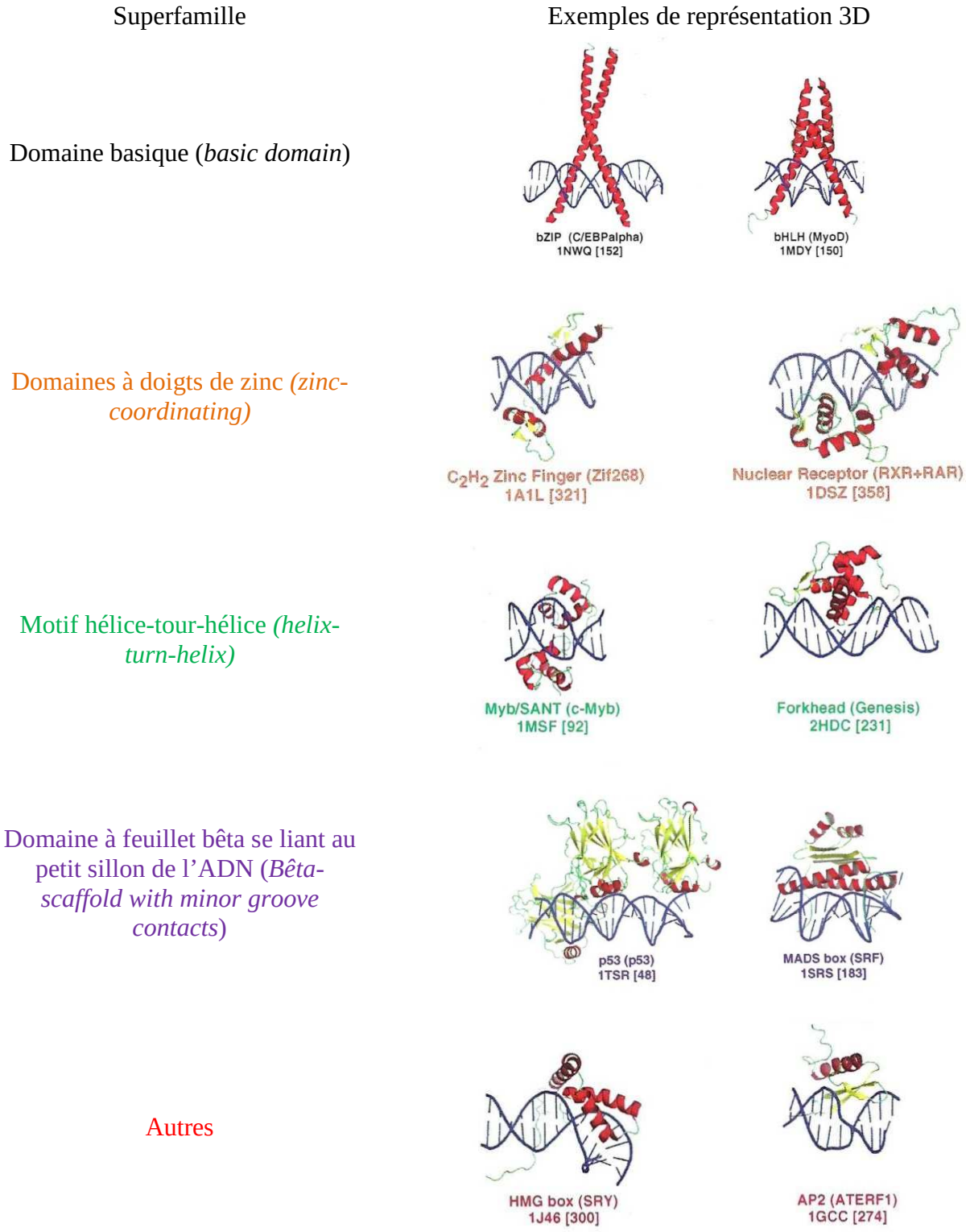
D. Les différentes familles de facteurs de transcription

Nous venons de voir que l'interaction d'un facteur de transcription avec l'ADN s'effectuait au niveau d'un domaine particulier, appelé le domaine de liaison à l'ADN. Ces domaines possèdent certains motifs particuliers qui reconnaissant une séquence spécifique d'ADN et vont venir s'intégrer parfaitement dans les grand et petit sillons du brin d'ADN (comme une voiture de train sur une voie de liaison), parce qu'ils exposent des acides aminés spécifiques aux endroits appropriés pour former des liaisons faibles avec les bases nucléotidiques. Des techniques de génétique moléculaire ont été utilisées pour modifier les acides aminés et tester si cela affecte l'affinité de liaison du facteur de transcription pour sa séquence cible (Phillips et Hoopes, 2008). La structure de l'ADN étant par nature extrêmement stable (double hélice), il était prévisible que la structure dans l'espace de la région de la protéine qui interagit avec l'ADN serait, elle aussi, relativement stéréotypée (Kaplan et Delpech, 2007).

C'est bien souvent à partir de ces motifs que les facteurs de transcription ont été classés et regroupés en superfamilles, familles, classes,… Cette classification a très souvent évolué au fur et à mesure des nouvelles découvertes et selon les auteurs, si bien qu'il est difficile parfois d'en faire ressortir le principal. Cependant, au travers des ouvrages et des publications, les nombreuses données accumulées semblent montrer qu'il n'existe qu'un nombre limité de structures protéiques susceptibles d'interagir avec l'ADN, qui ont été rangées en quatre superclasses (ou superfamilles) majeures, correspondant à quatre types principaux de structures : les domaines basiques (*basic domain*), les domaines à doigts de zinc (*zinc-coordinating*), les motifs hélice-tour-hélice (*helix-turn-helix*) et les domaines à feuillet bêta se liant au petit sillon de l'ADN (*bêta-scaffold with minor groove contacts*). Les autres facteurs qui ne rentrent pas dans ces catégories sont regroupés à part dans la classe « autres » (Figure 13). La proportion estimée de chacune de ces structures parmi tous les facteurs de transcription est présentée en Figure 14.

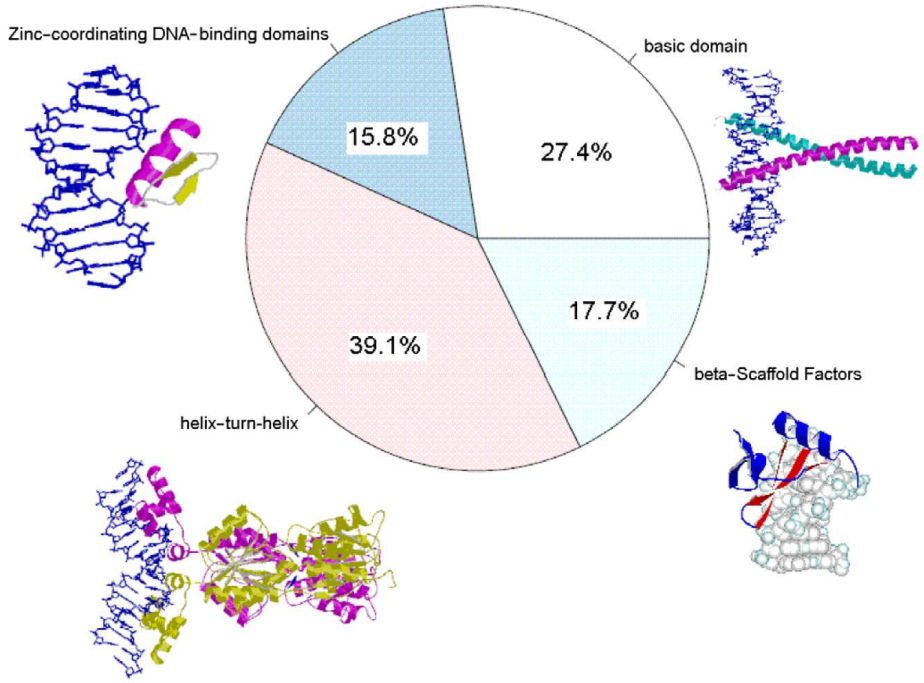
Les facteurs de transcription ont aussi parfois été classés selon leur fonction. Par exemple, chez de nombreux mammifères, y compris l'homme, un groupe important de gènes codant des facteurs de transcription et impliqués dans le développement des cellules, contiennent une séquence de 180 paires de base appelée *homeobox*. L'*homeobox* code un segment très conservé de 61 acides aminés appelé homéodomaine, qui reconnaît et se lie au promoteur des gènes cibles (Phillips et Hoopes, 2008). Ces protéines ont été regroupées sous le nom de protéines à homéodomaine ou homéoprotéines.

Figure 13 : Quelques exemples de représentations 3D d'interactions entre des facteurs de transcription et l'ADN



D'après (R. Hughes, 2011). Les structures ont été obtenues à partir du site PDB (*Protein Data Bank*, http://www.pdb.org/pdb/home/home.do). La famille du facteur de transcription est indiquée sous chaque structure, accompagné de son nom entre parenthèses, et de son numéro d'accès PDB en-dessous (le numéro entre crochets est un numéro de référence indiqué dans le livre). L'ADN est en bleu, les hélices alpha en rouge, les feuillets bêtas en jaune et les boucles en vert.

Figure 14 : Proportion des différentes familles de facteurs de transcription



D'après (Qian *et al.*, 2006). Les facteurs de transcription sont généralement classés en quatre catégories : les domaines basiques (*basic domain*, en haut à droite), les domaines à doigts de zinc (*zinc-coordinationg*, en haut à gauche), les motifs hélice-tour-hélice (*helix-turn-helix*, en bas à gauche) et les domaines à feuillet bêta se liant au petit sillon de l'ADN (*bêta-scaffold*, en bas à droite). Les structures 3D des complexes ADN-protéines ont été adaptées de (Luscombe *et al.*, 2000). Les proportions ont été calculées d'après la base de données TRANSFAC(v7.0) : http://www.gene-regulation.com/. La classe « autres » n'est pas incluse dans ce diagramme.

De nombreuses bases de données de facteurs de transcription existent maintenant sur internet : TFdb pour la souris (2004), TFCat pour l'homme et la souris (2009), TFCONES pour l'homme, la souris et le fugu (2007), ITFP pour l'homme, la souris et le rat (2008). La plupart contiennent également les séquences reconnues mais, comme le fait remarquer un article récent (Zhang *et al.*, 2012), elles se concentrent sur un ou deux génomes seulement. Les deux plus récentes bases de données sont :

- TRANSFAC® (Matys *et al.*, 2006) est une grande base de données de tous les facteurs de transcription chez les eucaryotes avec, pour chacun, la description du gène codant le facteur de transcription (séquence, position dans le génome, régulation) et des propriétés du facteur de transcription en lui-même : son site de fixation à l'ADN, sa classification (une classification a été faite selon le domaine de liaison à l'ADN), les gènes régulés,… Ce logiciel permet de rechercher un facteur de transcription, de faire des comparaisons, croiser les données,…

- AnimalTFDB (Zhang *et al.*, 2012) : c'est aujourd'hui la liste la plus complète des facteurs de transcription chez les animaux. Le site regroupe 72 familles de facteurs de transcription chez 50 espèces animales dont le génome a été séquencé. La base de donnée contient les caractéristiques des facteurs de transcription (structure, domaines fonctionnels, structure 3D, interaction protéique, domaine de liaison avec l'ADN, gènes-cibles,…) qui ont été triés par famille de domaine de liaison à l'ADN (selon TRANSFAC®) et par espèce. Le site donne par ailleurs les co-facteurs et les facteurs de remodelage la chromatine.

Il existe également une base de données de facteurs de transcription présumés (c'est-à-dire prévu comme tel par des études bioinformatiques mais non encore démontré expérimentalement) chez 930 génomes séquencés : DBD (*DNA-binding domain*, http://www.transcriptionfactor.org/index.cgi?Home).

Je vais présenter dans les deux paragraphes suivants quelques motifs de domaine de liaison (sur lesquels reposent les classifications actuelles) ainsi que quelques motifs de domaines d'activation de la transcription qui ont été bien décrits et caractérisés.

1. Les motifs des domaines de liaison à l'ADN

Parmi tous les motifs des domaines de liaison à l'ADN existants, quatre motifs particuliers ont été très bien décrits : le motif en doigt de gant (protéines dactyles) auquel est associé un atome de zinc, le motif hélice-tour-hélice, le motif hélice-boucle-hélice et la fermeture éclair à leucines (domaine bZIP). Les structures de ces différents motifs sont schématisées dans la Figure 15.

- Le motif hélice-tour-hélice (*helix-turn-helix*) : il est constitué de deux hélices alpha reliées entre elles par un coude bêta (Figure 15A). L'une des hélices interagit avec les bases de l'ADN au niveau du grand sillon, la seconde interagit avec l'enchaînement des désoxyriboses et des phosphates. Les facteurs transcriptionnels possédant ce motif semblent toujours agir sous forme de dimères. L'organisation des molécules est telle que les hélices homologues de chacun des monomères interagissent avec deux grands sillons successifs. Ce motif est retrouvé dans toutes les homéoprotéines (contenant un homéodomaine) mais aussi dans certaines autres protéines régulatrices impliquées dans le développement.

- Le motif dit en doigt de gant ou doigt de zinc (*zinc-finger*) : il est constitué d'une séquence d'une vingtaine d'acides aminés ayant dans l'espace une forme de doigt de gant (Figure 15B). Il en existe deux types : ceux qui contiennent quatre cystéines (Cys_4) et ceux qui contiennent deux cystéines et deux histidines (Cys_2/ His_2). Ces acides aminés qui définissent la nature du doigt sont situés à sa base (deux de chaque côté, les deux cystéines ou les deux histidines étant de chaque côté séparées par un à trois acides aminés). Un ion Zn^{++} est situé au centre du carré que forment ces quatre acides aminés : il y est fixé par des liaisons de coordination et est indispensable à l'activité de la protéine. Dans les motifs Cys_2/His_2 le remplacement des deux histidines par des cystéines abolit l'activité du facteur. Le motif en doigt de gant n'est jamais unique, on en retrouve au minimum deux, et jusqu'à plus d'une dizaine. Bien que le doigt interagisse avec les bases de l'ADN au niveau du grand sillon, il n'intervient que peu dans la spécificité de fixation du motif. Ce sont les acides aminés de la séquence polypeptidique qui les séparent (d'une longueur en général inférieure à 10 acides aminés) qui jouent le rôle le plus important dans la reconnaissance. Le facteur TFIIIA fut la première protéine à doigts de zinc découverte (elle en possède 9) : le nom « doigt de zinc » a initialement été donné pour décrire l'apparence semblable à des doigts d'un diagramme montrant la structure hypothétique de l'unité répétée dans le facteur de transcription IIIA de *Xenopus laevis* (Miller *et al.*, 1985). La protéine SP1 et les récepteurs nucléaires d'hormones stéroïdes et thyroïdiennes sont d'autres exemples de facteurs interagissant avec l'ADN grâce à ce type de motif.

- Le motif de type fermeture éclair à leucines (*leucine-zipper* ou bZIP) : les facteurs de transcription possédant ce motif sont toujours dimériques (Figure 15C), sous forme d'homodimères ou d'hétérodimères. Le monomère est constitué d'une séquence à caractère basique qui interagit de manière spécifique avec l'ADN (en formant une pince enchâssée dans le grand sillon) et d'un domaine hydrophobe en hélice alpha qui interagit avec le domaine homologue de l'autre chaîne. Dans ce domaine on retrouve une leucine tous les sept acides aminés, donc à chaque pas de l'hélice alpha. Toutes ces leucines se trouvent donc alignées, et c'est à leur niveau que se fait l'interaction

entre les deux monomères, d'où le nom de fermeture éclair à leucines. Les oncoprotéines Jun et Fos sont des exemples de facteurs interagissant avec l'ADN grâce à ce type de motif.

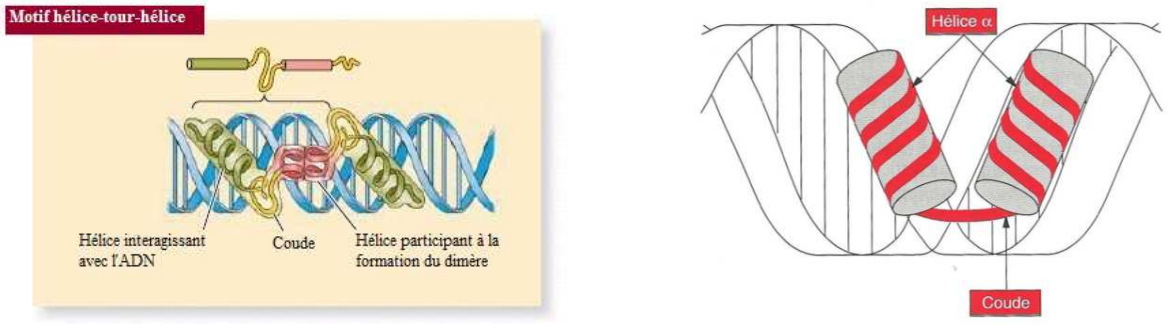
➢ Le motif hélice-boucle-hélice (*basic helix-loop-helix*, bHLH) : les facteurs de transcription possédant ce motif sont toujours dimériques (Figure 15D), sous forme d'homodimères ou d'hétérodimères (même si le plus souvent seuls les hétérodimères sont fonctionnels). En allant de l'extrémité NH_2 vers l'extrémité COOH, il est constitué d'un domaine riche en acides aminés basiques (qui interagit avec l'ADN et assure la spécificité du facteur), d'un domaine en hélice alpha (qui interagit par des liaisons hydrophobes avec l'hélice homologue de l'autre oligomère du dimère), d'une boucle qui est libre et d'un second domaine en hélice alpha (qui, comme le premier, interagit par le même type de liaisons avec son homologue de l'autre oligomère). La séquence de la boucle n'est pas indifférente et joue un rôle dans la spécificité. Les protéines E12, E47 et MyoD sont des exemples de facteurs interagissant avec l'ADN grâce à ce type de domaines. La protéine MyoD (impliquée dans la détermination des myoblastes) en particulier se lie aux promoteurs de gènes responsables de la différenciation musculaire. MyoD se lie aussi à son propre promoteur, maintenant ainsi ses propres niveaux d'expression dans les cellules musculaires (Phillips et Hoopes, 2008).

Les deux derniers motifs sont désormais regroupés au sein de la superfamille des domaines basiques (*basic domains*).

Figure 15 : Structure de quatre motifs de domaines de liaison à l'ADN des facteurs de transcription chez les eucaryotes

D'après (Phillips et Hoopes, 2008) | D'après (Kaplan et Delpech, 2007)

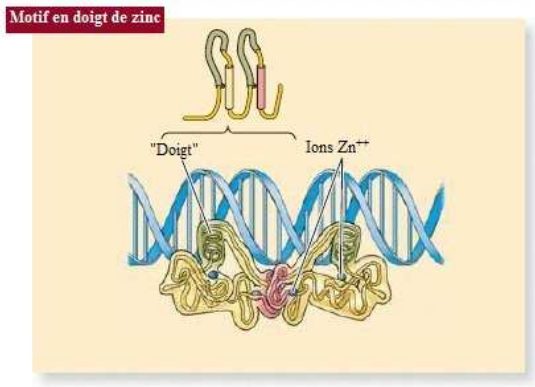
A. Structure du motif hélice-tour-hélice (*helix-turn-helix*)

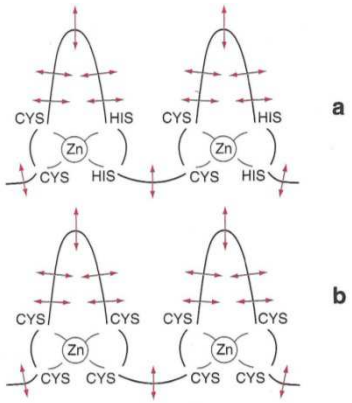


Exemples (sites de liaison) : Homéoprotéines (promoteurs de gènes du développement) , Pit-1 (promoteurs des gènes de la GH, de la TSH, et de la prolactine), Oct-1, Oct-2 (promoteurs des gènes d'immunoglobulines, de l'histone H2B et de snRNA).

Figure 15 (suite)

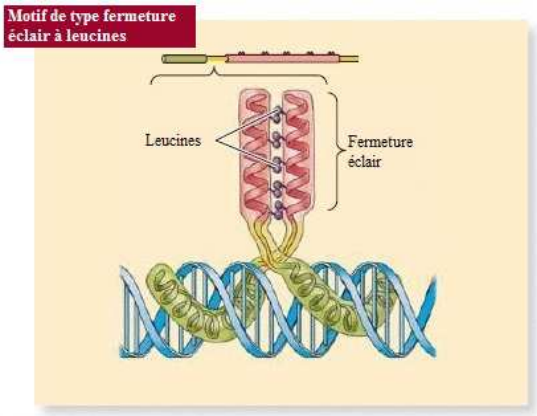
B. Structure du motif dit en doigt de gant ou doigt de zinc (*zinc-finger*)

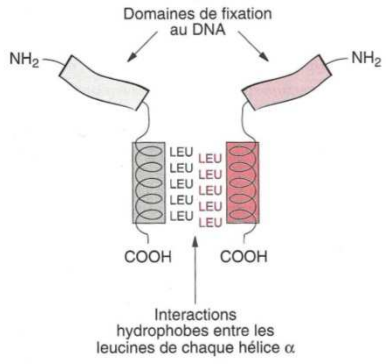




Il existe deux classes différant l'une de l'autre par les acides aminés liant le zinc : en (a), le motif CYS2/HIS2, et en (b) le motif CYS4.
Exemples (sites de liaison) : TFIIIA (promoteur des gènes des ARN 5S du ribosome), récepteurs nucléaires d'hormones (promoteurs régulés par les hormones thyroïdiennes, stéroïdes, par l'acide rétinoïque,...), SP1 (séquence GGGCGG, promoteur de nombreux gènes de ménage ou *house keeping*).

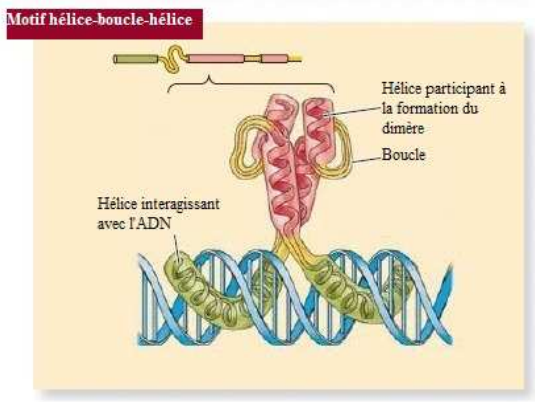
C. Structure du motif de type fermeture éclair à leucines (*leucine-zipper* ou bZIP)

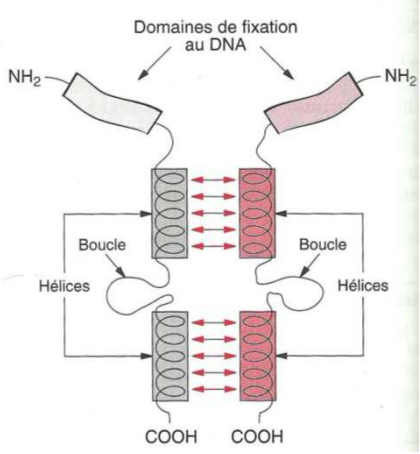




Exemples (sites de liaison) : Fos/Jun (site API : TGACTGNCAA), C/EBP (séquence stimulatrice de transcription CCAAT), CREB (séquence TGACGTCA, promoteurs de gènes contrôlés par l'AMPc).

D. Structure du motif hélice-boucle-hélice (*helix-loop-helix*)





Exemple (sites de liaison) : MyoD (promoteurs de gènes exprimés spécifiquement dans le muscle).

2. Les motifs des domaines d'activation de la transcription

Le second domaine indispensable des facteurs transcriptionnels est celui qui agit sur la transcription. Comme pour les domaines de fixation à l'ADN, le nombre de motifs de base semble limité, certains facteurs peuvent en posséder plusieurs. Bien qu'une classification stricte soit difficile à réaliser, schématiquement trois principaux types de motifs ont été caractérisés, mais il peut en exister d'autres. Il s'agit (Kaplan et Delpech, 2007) :

- du domaine riche en acides aminés acides dont le modèle est le facteur de levure Gal4.

- du domaine riche en glutamines (25 %), dont le modèle est le facteur de transcription SP1.

- du domaine riche en prolines (20 à 30 %), dont le modèle est le facteur de fixation à la boîte CAAT, NF1/CTF.

E. Les facteurs de transcription : moteurs de l'évolution et de la complexification des individus eucaryotes

Plusieurs généticiens ont constaté que la complexité d'un organisme (diversité des types cellulaires) n'était pas corrélée au nombre de gènes. En effet, de nombreux vertébrés ont seulement environ deux fois plus de gènes que les invertébrés (Tableau 1), et beaucoup d'entre eux sont le résultat de duplication des gènes existants, plutôt que de la création de nouveaux gènes (Levine et Tjian, 2003). Ainsi, les scientifiques se sont demandés : comment, au cours de l'évolution, les organismes sont-ils devenus d'autant plus complexes et diversifiés sans avoir beaucoup plus de gènes ?

Une des réponses à cette question pourrait se trouver dans le développement d'une façon plus complexe de contrôler l'expression des gènes (signalisation cellulaire, épissage alternatif de l'ARN, réarrangements de l'ADN, siARN, condensation ou autres modifications de la chromatine,…), et notamment la diversification des facteurs de transcription. Ainsi, la complexité d'un organisme (diversité des types cellulaires) n'est pas corrélée au nombre de gènes, mais pourrait être corrélée à la complexité de la régulation de l'expression de ses gènes et au ratio de facteur de transcription. Une grande partie de la complexité de la différenciation des cellules animales et végétales, peut être attribuée à l'évolution des systèmes complexes constitués de séquences courtes (6 à 8 paires de bases) cis-régulatrices ADN ou des motifs, ainsi que les facteurs de transcription qui se lient aux motifs et interagissent les uns avec les autres. Les quelques données pour appuyer cette hypothèse sont (Levine et Tjian, 2003) :

- la complexité du contrôle transcriptionnel peut être illustrée en comparant le nombre et la localisation d'éléments cis chez les eucaryotes supérieurs et inférieurs. Par exemple, la drosophile a généralement plusieurs *enhancer* pour un seul gène de 2 à 3 kilobases, éparpillés sur une vaste région (10 kilobases) de l'ADN, alors que la levure n'a pas d'*enhancers*. La régulation à long terme semble être indicative de la nécessité d'un niveau élevé de contrôle sur les gènes impliqués dans le développement et la différenciation cellulaire. Des promoteurs plus complexes sont donc corrélés avec une complexification des organismes au cours de l'évolution. La complexité peut reposer sur la combinaison de multiples *enhancers, silencers* et promoteurs différents pour réguler une unité transcriptionnelle (exemple : expression sélective des gènes des immunoglobulines dans les lymphocytes B).

- Il existe des TAF (*TBP Associated Factor*, voir I.B) et des TRF (*TBP-related factors*) tissus-spécifiques, présents dans les organismes plus évolués. L'énorme palette d'expression génique serait le résultat de la combinaison entre des TAF spécifiques avec de multiples *enhancers*.

- Les cofacteurs et les enzymes modifiant ou remodelant la chromatine ont rapidement évolué au cours de l'évolution.

- L'évolution des familles de facteurs de transcription : les organismes supérieurs disposent d'un grand nombre de familles diverses de facteur de transcription définis par la séquence de leurs domaines de liaison à l'ADN. Les études évolutives ont montré que, bien que le motif de liaison à l'ADN soit hautement conservé parmi les plantes et les animaux, le reste des séquences protéiques de ces organismes est souvent très différente. En outre, une famille de facteurs de transcription particuliers peut avoir des rôles différents chez les plantes et chez les animaux, et quelques nouveaux facteurs de transcription ont évolué dans chaque règne depuis leur divergence (Phillips et Hoopes, 2008). De plus, de nombreux facteurs de transcription au sein de la même famille travaillent souvent ensemble pour réguler la transcription d'un seul gène ce qui augmente encore plus le niveau de complexité génétique des eucaryotes.

- Le génome de la levure code environ 300 facteur de transcription, soit un pour 20 gènes, tandis que celui de l'homme code environ 3 000 facteur de transcription, soit un pour 10 gènes. Avec le contrôle combinatoire, la double augmentation du « taux » de facteurs de transcription par gène se traduit en fait par beaucoup plus de combinaisons possibles d'interactions, permettant l'augmentation spectaculaire de la diversité des organismes. Lorsque nous considérons les complexités supplémentaires de remodelage de la chromatine, de la régulation de la stabilité de l'ARNm, et du contrôle de la traduction, il est facile de comprendre comment les cellules des organismes supérieurs peuvent produire une énorme variété de réponses génétiques aux signaux environnementaux.

Compte tenu de la fonction des facteurs de transcription, et lorsque nous considérons les complexités supplémentaires de remodelage de la chromatine, de la régulation de la stabilité de l'ARNm, et du contrôle de la traduction, il devient plus facile de comprendre comment les cellules des organismes supérieurs sont capables de produire une énorme variété de réponses génétiques aux signaux environnementaux avec si peu de gènes, et d'adapter leur réponse à un niveau de contrôle fin sur l'ADN en permettant des gradations d'expression.

Tableau 1 : Comparaison des génomes de quelques organismes

Organisme	Taille du génome en Mb	Estimation du nombre de gènes	Estimation du nombre de facteurs de transcription	Compacité (nombre de kb pour un gène)*	Nombre de gènes/nombre de facteurs de transcription
Virus	1 à 100 kb	4 à 10	0 à 1	1 à 5	-
Escherichia coli (eubactérie)	4,6 Mb	4 000	270	1	1 facteur de transcription pour 15 gènes
Saccharomyces cerevisiae (levure)	12,1 Mb	6 000	300	2	1 facteur de transcription pour 20 gènes
Caenorhabditis elegans (invertébré nématode)	97 Mb	19 000	1 000	5	1 facteur de transcription pour 19 gènes
Drosophila melanogaster (invertébré insecte)	137 Mb	14 000	1 000	10	1 facteur de transcription pour 14 gènes
Fugu rubripes (vertébré poisson)	365 Mb	22 000	1 274	11	1 facteur de transcription pour 17 gènes
Mus musculus (mammifère rongeur)	2 500 Mb	30 000	3 000	100	1 facteur de transcription pour 10 gènes
Homo sapiens sapiens (mammifère primate)	2 900 Mb	30 000	3 000	100	1 facteur de transcription pour 10 gènes
Arabidopsis thaliana (plante herbacée dicotylédone)	125 Mb	25 000	2 296	2	1 facteur de transcription pour 10 gènes

* Les valeurs données correspondent à des tailles moyennes de gènes, mais n'impliquent pas que la distribution des gènes soit uniforme.

D'après (Levine et Tjian, 2003), (Kaplan et Delpech, 2007) et la base de données AnimalTFDB (http://www.bioguo.org/AnimalTFDB/family_index.php). Chez l'homme la molécule d'ADN par génome haploïde (3 milliards de paires de bases) représenterait une longueur physique de 1 mètre entièrement déroulé.

II. Maladies associées à des facteurs de transcription chez les mammifères

Les facteurs de transcription jouant un rôle essentiel dans l'expression des gènes et étant directement impliqués dans le développement, la signalisation intercellulaire et le cycle cellulaire, on comprend que des mutations ou des dérèglements de ces protéines sont souvent associées à diverses maladies, allant du cancer (défini comme une maladie génétique), aux maladies auto-immunes, en passant par divers troubles du développement. On en déduit que les organismes avec une mutation sévère dans un gène codant un facteur de transcription présentent des irrégularités profondes dans l'organisation et le développement, pouvant même être létales à court ou à long terme (Phillips et Hoopes, 2008).

Ce paragraphe ne prétend pas être un catalogue de toutes les mutations connues des facteurs de transcription chez les mammifères : le but est de montrer l'importance et les conséquences fonctionnelles de ces mutations en se focalisant sur quelques exemples connus et bien décrits. C'est en général plus la cause d'une perte de fonction que d'une surexpression.

Pour réaliser un panorama des mutations de facteurs de transcription chez les mammifères, j'ai utilisé trois bases de données : OMIA (http://omia.angis.org.au/home/), OMIM (http://www.omim.org/) et MGI (http://www.informatics.jax.org/). La base OMIA (*Online Mendelian Inheritance in Animals*) est une base de données en ligne initialement rédigée par le professeur Frank Nicholas de l'Université de Sydney (Australie), qui recense les maladies héréditaires et les gènes associés de 212 espèces animales autres que le rat, la souris et l'homme (qui ont leurs propres ressources), et qui contient des liens vers les bases NCBI (http://www.ncbi.nlm.nih.gov.gate2.inist.fr/), OMIM (http://www.omim.org/) et Ensembl (http://www.ensembl.org/index.html). OMIM (*Online Mendelian Inheritance in Man*) est une base de données originellement compilée par Victor Almon McKusick, qui dresse un catalogue de toutes les maladies héréditaires connues et, si possible, les relie aux gènes adéquats au sein du génome humain. MGI (*Mouse Genome Informatics*) est une base de données en ligne organisée par le Jackson Laboratory, qui regroupe une grande quantité d'informations importantes sur la génétique, la génomique et la biologie de la souris de laboratoire : elle intègre un catalogue des souris modèles pour les maladies humaines, avec des liens vers OMIM et NCBI. En utilisant ces trois immenses bases de données, je me suis focalisé sur les maladies ayant été clairement associées à des facteurs de transcription bien caractérisés. Chez les espèces de mammifères autres que l'homme et la souris, six gènes ont été décrits et la mutation responsable a été trouvée : c'est ceux que je vais développer par la suite. Chez l'homme et la souris, les résultats étant trop nombreux, je me suis arrêté aux gènes qui ont servi pour créer des modèles de souris pour la maladie humaine correspondante. La mutation responsable n'est pas toujours connue (il peut y en avoir plusieurs chez l'homme alors qu'il n'y en a généralement qu'une chez les autres espèces de mammifères). Les 85 résultats de cette recherche figurent dans le tableau en Annexe 2.

A. DMRT3 (*Doublesex- and Mab-3-Related Transcription factor 3-like*)

Rôle : facteur de transcription (actuellement répertorié comme LOC100147177 dans NCBI) avec un domaine DM, domaine qui a été découvert dans les protéines *doublesex* de *Drosophila melanogaster* (d'où l'abréviation DM), chez qui il agit sur des gènes de la détermination du sexe, sur la différenciation des neuroblastes et sur des gènes codant des protéines du sac vitellin. L'analyse détaillée des souris KO pour *Dmrt3* a révélé que DMRT3 a un rôle central pour la configuration des circuits spinaux contrôlant les allures chez les vertébrés. Ce facteur jouerait donc un rôle dans la configuration des circuits spinaux contrôlant l'allure chez les vertébrés. Il serait impliqué dans la spécification neuronale à l'intérieur de la moelle épinière et dans le développement d'un réseau de locomotion coordonnée contrôlant les mouvements des membres (Andersson *et al.*, 2012). Il pourrait réguler la transcription au cours du développement sexuel.

Maladie associée : « gaitedness » chez *Equus caballus* (cheval).
Mode de transmission : caractère multifactoriel.
Races touchées : les races « gaited » (ayant cette capacité) sont entre autres le cheval Islandais, le Kentucky Mountain Saddle Horse, le Missouri Fox Trotter, le Paso Fino, le Paso péruvien, le Rocky Mountain Horse et le Tennessee Walking Horse. Les races « non-gaited » comprennent le cheval Arabe, le poney Gotland, le cheval de trait Suédois du Nord, le cheval de Przewalski, le poney Shetland, le Swedish Ardennes, le Suédois (demi-sang et pur-sang).
Découverte : étude d'association avec des SNP (Andersson *et al.*, 2012).
Mutation décrite : mutation non-sens (Ser301STOP) dans le gène *Dmrt3* qui produit une protéine raccourcie.
Phénotype : voir Figure 16.
Maladie homologue humaine : aucune.

Figure 16 : « Gaitedness » chez le cheval

D'après (Andersson *et al.*, 2012). Les chevaux ont trois allures naturelles qui sont, par ordre de vitesse croissante, le pas (quatre temps, latéral), le trot (deux temps, en diagonale) et le galop (quatre temps). Certaines races de chevaux sont capables en plus d'effectuer d'autres allures particulières à des vitesses intermédiaires : l'amble (une allure à deux temps) ou le tölt (une allure à quatre temps) en sont des exemples. Certaines races sont capables d'effectuer quatre allures (par exemple pas, tölt, trot et galop) et d'autres cinq (par exemple pas, tölt, trot, galop et amble). La capacité de produire ces allures particulières a été nommée par les Anglais « gaitedness » (*gait* signifie allure en Anglais). Le cheval Islandais fait partie des races « gaited », c'est-à-dire ayant cette capacité (sur la photo de gauche on peut voir un cheval Islandais à l'amble) alors que le cheval Arabe fait partie des races « non-gaited » (Andersson *et al.*, 2012). Chez les chevaux, une mutation dans le gène *Dmrt3* permet d'effectuer des changements d'allures et a un effet favorable sur les performances des courses attelées (Petersen *et al.*, 2013). Chez les chevaux Islandais, l'homozygotie pour la mutation non-sens dans le gène *Dmrt3* est nécessaire (mais pas suffisante) pour effectuer l'amble dans cette race (sur la photo de droite on peut voir un cheval Islandais au trot).

B. FOXI3 (*Forkhead box I3*)

Rôle : facteur de transcription, auparavant inconnu, qui contribuerait au développement de structures ectodermiques.
Maladie associée : dysplasie ectodermique canine (*canine ectodermal dysplasia*, CED) chez *Canis lupus familiaris* (chien).
Mode de transmission : caractère mendélien, autosomique dominant.
Races touchées : c'est un caractère désiré dans plusieurs races, dont le chien nu du Pérou, le chien nu du Mexique (Xoloitzcuintli), et le chien Chinois à crête. Il existe deux variétés dans chaque race : une variété nue et une variété poilue ou à fourrure (appelée « houpette à poudre » pour le chien Chinois à crête). La variété nue présente une hypotrichose et des anomalies dentaires.
Découverte : étude d'association avec des SNP (Drögemüller *et al.*, 2008).
Mutation décrite : duplication de 7 pb dans l'exon 1 du gène *FOXI3* qui produit un décalage du cadre de lecture entraînant l'apparition d'un codon stop prématuré (Drögemüller *et al.*, 2008).

Phénotype : voir Figure 17.
Maladie homologue humaine : Dysplasie ectodermique hypohidrotique (DEH) liée à l'X (DEX) (*X-linked hypohidrotic ectodermal dysplasia,* XHED).

Figure 17 : Dysplasie ectodermique chez le chien nu

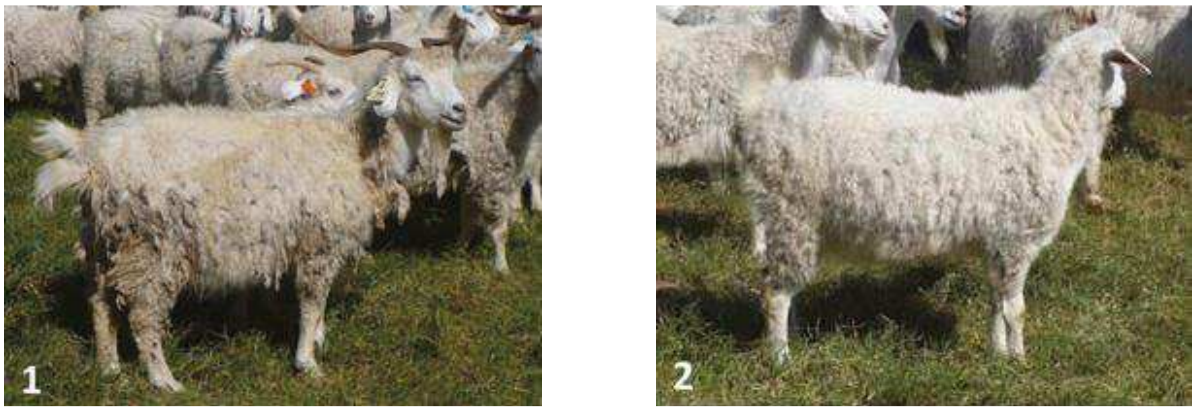
Race	Chien chinois à crête	Chien nu du Pérou	Chien nu du Mexique (Xoloitzcuintli)
Variété nue			
Variété poilue			

Photos issues du site du Club du Chihuahua, du Coton de Tuléar et des Exotiques (CCCE : http://www.ccce.org/races.html). Les chiens nus ont des degrés de perte de poils variable mais le plus souvent presque tout le corps est affecté. En fonction de la race, il y aura des touffes de poils sur les extrémités des membres et le dessus de la tête. Certains chiens ont des dents très anormales, dont beaucoup sont absentes. D'autres semblent n'avoir que de légères anomalies. Les ongles peuvent être longs et cassants. Les décès néonataux sont plus fréquents chez les chiots nus (l'homozygotie pour la mutation est létale) que chez les chiots avec poils. Des auteurs (Wiener *et al.*, 2013) ont décrit les différences cliniques et histologiques entre les trois sous-phénotypes (vrais nus, demi-poilus et « houpette à poudre ») de chiens Chinois à crête et les différences distinctes entre la dysplasie ectodermique canine chez les chiens Chinois à crête et la dysplasie ectodermique canine liée à l'X (autre syndrome). L'examen histologique de la peau glabre et des coussinets a montré une absence de follicules pileux, de structures annexes, et de glandes eccrines.

C. FOXL2 (*Forkhead box L2*)

Rôle : facteur de transcription appartenant à la famille des FOX pouvant jouer un rôle dans le développement et la fonction ovarienne.
Maladie associée : *Polled/Intersex syndrome* (PIS) : association de l'absence de cornes (une chèvre sans corne est appelée une chèvre motte) et d'intersexualité chez *Capra hircus* (chèvre).
Mode de transmission : caractère mendélien, autosomique dominant. Comme dans d'autres espèces, l'absence de cornes chez les caprins est un caractère dominant (P, *polled*) dû à un unique locus autosomique. Mais contrairement à d'autres espèces, ce même allèle est autosomique récessif pour l'intersexualité. Quelques chèvres XY qui sont homozygotes pour l'allèle *P*, sont stériles. La plupart des chèvres XX qui sont homozygotes pour l'allèle *P*, montrent des signes d'intersexualité, allant de la femelle presque normale à un phénotype mâle, avec développement des tractus et des organes reproducteurs.
Race touchée : a été d'abord décrit chez la Saanen, mais existe aussi dans d'autre race (chèvre cachemire et chèvres Boer notamment).
Découverte : étude d'association avec des microsatellites, confirmé par SNP (Kijas *et al.*, 2013).
Mutation décrite : le locus *PIS* a été localisé dans la région télomérique du chromosome 1 caprin et a été associé à un gène que l'on a appelé *PISRT1* (*PIS-regulated transcript 1*, qui code un ARN non-codant de 1 500 pb). On a pu montrer que le phénotype anormal est causé par une délétion de 11,7 kb (Pailhoux *et al.*, 2001) qui affecte la transcription de *PISRT1* et d'un gène voisin, *FOXL2*, bien que localisés respectivement à 20 kb et 200 kb de la délétion. La zone délétée ne contient à ce jour aucune séquence codante identifiée, mais 80 % de séquences répétées de type LINE. Cette région pourrait constituer un « interrupteur » transcriptionnel dépendant d'une protéine clef du déterminisme du sexe, telle SRY (*sex determining region Y*), qui affecterait à distance la régulation transcriptionnelle des gènes de la région (Pailhoux *et al.*, 2005).
Phénotype : voir Figure 18.
Maladie homologue humaine : syndrome blépharophimosis-ptosis-épicanthus inversus (*Blapharophimosis-Ptosis-Epicanthus inversus Syndrome*, BPES), qui entraîne chez l'homme une malformation des paupières associée à un dysfonctionnement ovarien (ménopause précoce) chez des patientes à caryotype normal 46, XX. Cette maladie est due à une mutation dans le gène *FOXL2*.

Figure 18 : Syndrome *Polled/Intersex* (PIS) chez la chèvre

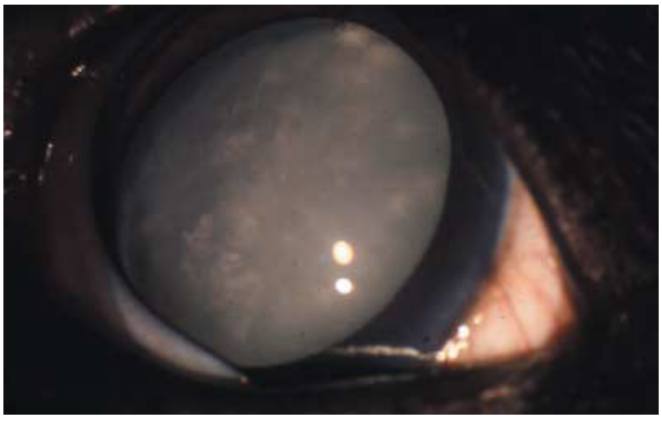


D'après (Kijas *et al.*, 2013). Chèvres cachemire avec cornes (1) ou sans corne (2). Pour des raisons zootechniques, l'absence de cornage a été recherchée chez les Caprins. Mais chez cette espèce, cette sélection se heurte à un obstacle majeur : l'anomalie est toujours associée à l'inversion sexuelle des femelles. Il n'a jamais été observé de recombinaison entre le gène d'absence de corne, qui se transmet de façon dominante, et celui de l'intersexualité, dont la transmission est récessive. C'est cette association qui a poussé à rechercher un gène autosomique de différenciation sexuelle en prenant la chèvre comme modèle d'étude. Ce modèle, chèvre sans corne ou motte (*Polled* (P) pour les Anglo-saxons), est en effet idéal puisque la maladie de la différenciation sexuelle est liée à un marqueur phénotypique apparent.

D. HSF4 (*Heat Shock transcription factor 4*)

Rôle : les facteurs de transcription de choc thermique (HSF, *heat shock transcription factors*) activent les gènes de réponse au choc thermique dans des conditions de chaleur ou autres stress. HSF4 n'a pas la répétition de la partie hydrophobe carboxy-terminale qui est partagée entre tous les facteurs de transcription de cette famille chez les vertébrés. Ce facteur semble impliqué dans la régulation négative de l'activité de liaison à l'ADN.
Maladie associée : cataracte héréditaire (PHC, *Primary hereditary cataract*) chez *Canis lupus familiaris* (chien).
Mode de transmission : autosomique dominant ou récessif selon la race.
Races touchées : Berger Australien (transmission autosomique dominante), Boston Terrier et Staffordshire Bull Terrier (transmission autosomique récessive).
Découverte : étude d'association avec des SNP (Mellersh *et al.*, 2006).
Mutation décrite : chez le Berger Australien, une délétion d'un nucléotide C à la même position de l'exon 9 (g.85286582delC), qui est censée modifier le cadre de lecture et introduire un codon stop prématuré 177 nucléotides (59 aminé aminés) en aval de la délétion (Mellersh *et al.*, 2009). Chez le Boston Terrier et le Staffordshire Bull Terrier, insertion d'une cytidine dans l'exon 9 (CFA5 g85286582 - 85286583insC) qui modifie le cadre de lecture du gène et introduit un codon stop prématuré (Mellersh *et al.*, 2006).
Phénotype : voir Figure 20
Maladie homologue humaine : cataracte de Marner.
Phénotype : voir Figure 19.
Maladie homologue humaine : aucune

Figure 19 : Cataracte héréditaire chez un Staffordshire Bull Terrier



D'après (Mellersh *et al.*, 2006). Cataracte chez un Staffordshire Bull Terrier de 3 ans et demi. L'image a été prise après mydriase.

E. LHX3 (*LIM homeobox 3*)

Rôle : facteur de transcription possédant un domaine LIM (nommé ainsi d'après les premières protéines découvertes Lin11, Isl-1 et Mec-3), essentiel pour la formation de la glande pituitaire (hypophyse) et dans le développement des motoneurones.
Maladie associée : nanisme hypophysaire (*Combined Pituitary Hormone Deficiency*, CPHD) chez *Canis lupus familiaris* (chien)
Mode de transmission : caractère mendélien, autosomique récessif.
Race touchée : Berger Allemand

Découverte : étude d'association avec des marqueurs microsatellites (Voorbij *et al.*, 2011), confirmés par SNP (Tsai *et al.*, 2012).
Mutation décrite : délétion de l'une des six répétitions de 7 pb dans l'intron 5 de *LHX3* (chromosome 9), réduisant la taille de celui-ci à 68 pb et provoquant un défaut d'épissage et l'apparition d'un résidu asparagine dans l'homéodomaine de LHX3 (Voorbij *et al.*, 2011).
Phénotype : voir Figure 20.
Maladie homologue humaine : déficit hypophysaire combiné multiple (*Combined Pituitary Hormone Deficiency*, CPHD)

Figure 20 : Nanisme hypophysaire (*Combined Pituitary Hormone Deficiency*, CPHD) chez le chien Berger Allemand

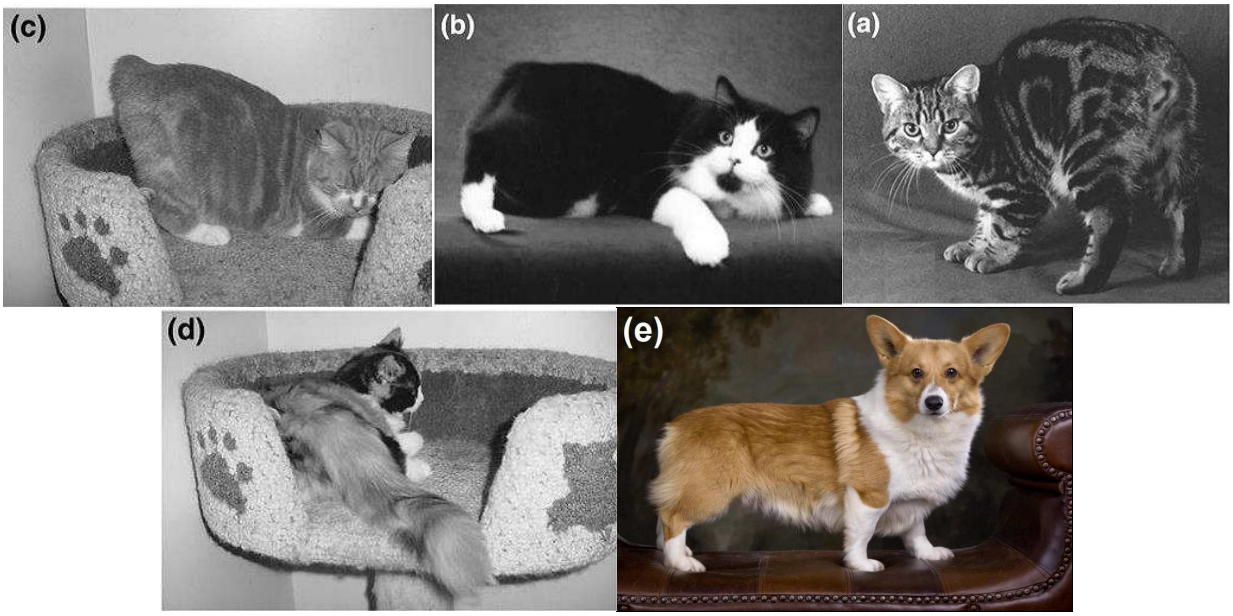
D'après (Voorbij *et al.*, 2011). Deux mâles bergers allemands de la même portée, âgés de quatorze mois : celui de gauche est en bonne santé, celui de droite (son frère) est touché par le nanisme hypophysaire. Notez le retard de croissance proportionnel, la persistance des poils de chiot et le manque de poils de garde (poils de couverture) du nain.

F. T (*tail*) ou brachyury

Rôle : facteur de transcription embryonnaire appartenant à la famille T-*box* (région N-terminale qui se lie à un élément spécifique de l'ADN, le site palindromique T) affectant la transcription des gènes nécessaires à la formation de mésoderme et la différenciation. La protéine est localisée dans les cellules dérivées de la notochorde.
Maladie associée : queue courte ou bob-tail (queue écourtée en Anglais) ou mutation brachyury chez *Canis lupus familiaris* (chien) et chez *Felis catus* (chat). Existe aussi chez les bovins et le mouton mais le gène responsable n'est pas identifié. C'est pour cette raison que le gène a été nommé T (pour « tail ») ou brachyury (du grec « brakhus » signifiant court et « oura », queue).
Mode de transmission : caractère mendélien, autosomique dominant (létal très tôt dans le développement embryonnaire si homozygote)
Races touchées chez *Canis lupus familiaris* (chien) : Welsh Corgi Pembroke, Berger Australien, par exemple.
Découverte chez *Canis lupus familiaris* (chien) : séquençage du gène T chez plusieurs races de chiens (Haworth *et al.*, 2001).
Mutation décrite chez *Canis lupus familiaris* (chien) : une mutation faux-sens (C295G; Ile63Met) dans le gène T (Haworth *et al.*, 2001) est située dans une région hautement conservée du domaine T-*box* et modifie la capacité de la protéine à se lier à sa séquence-cible. Une enquête ultérieure sur

cette mutation (maintenant appelé C189G) dans d'autres races de chiens a montré que dans les 17 races chez lesquelles la mutation C189G a été observée, il y avait une corrélation parfaite entre cette mutation et le phénotype à queue courte. Cependant, certaines races à queue courte ne portent pas cette mutation ni d'autres mutation dans le gène T (Hytönen *et al.*, 2009). D'autres facteurs génétiques affectant la longueur de la queue sont donc encore à découvrir.
Races touchées chez *Felis catus* (chat) : Bobtail américain, Manx et Pixie-Bob.
Découverte chez Felis catus (chat) : séquençage du gène T dans plusieurs lignées indépendantes de chats Manx (Buckingham *et al.*, 2013).
Mutation décrite chez *Felis catus* (chat) : trois délétions de 1 pb [c.998delT; c.1169delC; c.1199delC] et une duplication/délétion [c.998_1014dup17delGCC], chacune provoquant un décalage du cadre de lecture qui mène à l'apparition d'un codon stop prématuré et à la troncature de l'extrémité carboxy-terminale de la protéine (Buckingham *et al.*, 2013).
Phénotypes : voir Figure 21.
Maladie homologue humaine : sensibilité aux anomalies du tube neural.

Figure 21 : Mutations brachyury chez le chien et le chat



Depuis qu'il avait été remarqué que les souris invalidées pour le gène *T* avaient un phénotype très semblable à celui des chiens et des chats à queue courte, l'homologue canin et félin du gène *T* a été vu comme un gène candidat pour expliquer la queue courte chez les chats Bobtail américain, Manx et Pixie-Bob (Buckingham *et al.*, 2013) et chez les chiens Welsh Corgis Pembroke (Haworth *et al.*, 2001). La longueur variable de la queue des Manx peut être classée en quatre phénotypes distincts (Buckingham *et al.*, 2013) : absence complète de la queue ou anourie (chat de variété *rumpy*, photo a), queue minimale détectable seulement par palpation (variété *rumpy-riser*, photo b), queue courte (variété *stumpy*, photo c), et queue complète (variété *longie*, photo d). La longueur de la queue est proportionnelle au nombre de vertèbres caudales, avec absence complète de vertèbres caudales constaté uniquement chez les chats avec le phénotype *rumpy*. La photo e (issue du site infoVeto, http://www.infoveto.com/race/welsh-corgi/) présente un chien Welsh Corgi Pembroke complètement anoure.

III. Applications potentielles des connaissances concernant les facteurs de transcription mammaliens

A. La reprogrammation de cellules matures

D'après le communiqué de presse du 10 août 2012 du site officiel du Prix Nobel ("The 2012 Nobel Prize in Physiology or Medicine - Press Release," 2012).

Chaque individu est issu d'un ovocyte fécondé. Pendant les premiers jours qui suivent la conception, l'embryon est constitué de cellules immatures, appelées cellules souches pluripotentes car chacune est capable de se développer dans tous les types cellulaires qui forment l'organisme adulte. Au cours du développement embryonnaire, ces cellules se spécialisent et donnent naissance à un type de cellules particulier (cellules nerveuses par exemple). On pensait que ce voyage d'une cellule immature vers une cellule spécialisée était unidirectionnel, que les modifications au cours de la maturation ne permettaient plus de revenir à un stade pluripotent immature.

Le Prix Nobel de physiologie ou médecine 2012 a été attribué conjointement à Sir John Bertrand Gurdon et à Shinya Yamanaka, deux scientifiques qui ont découvert que des cellules spécialisées matures pouvaient être reprogrammées pour devenir des cellules immatures pluripotentes.

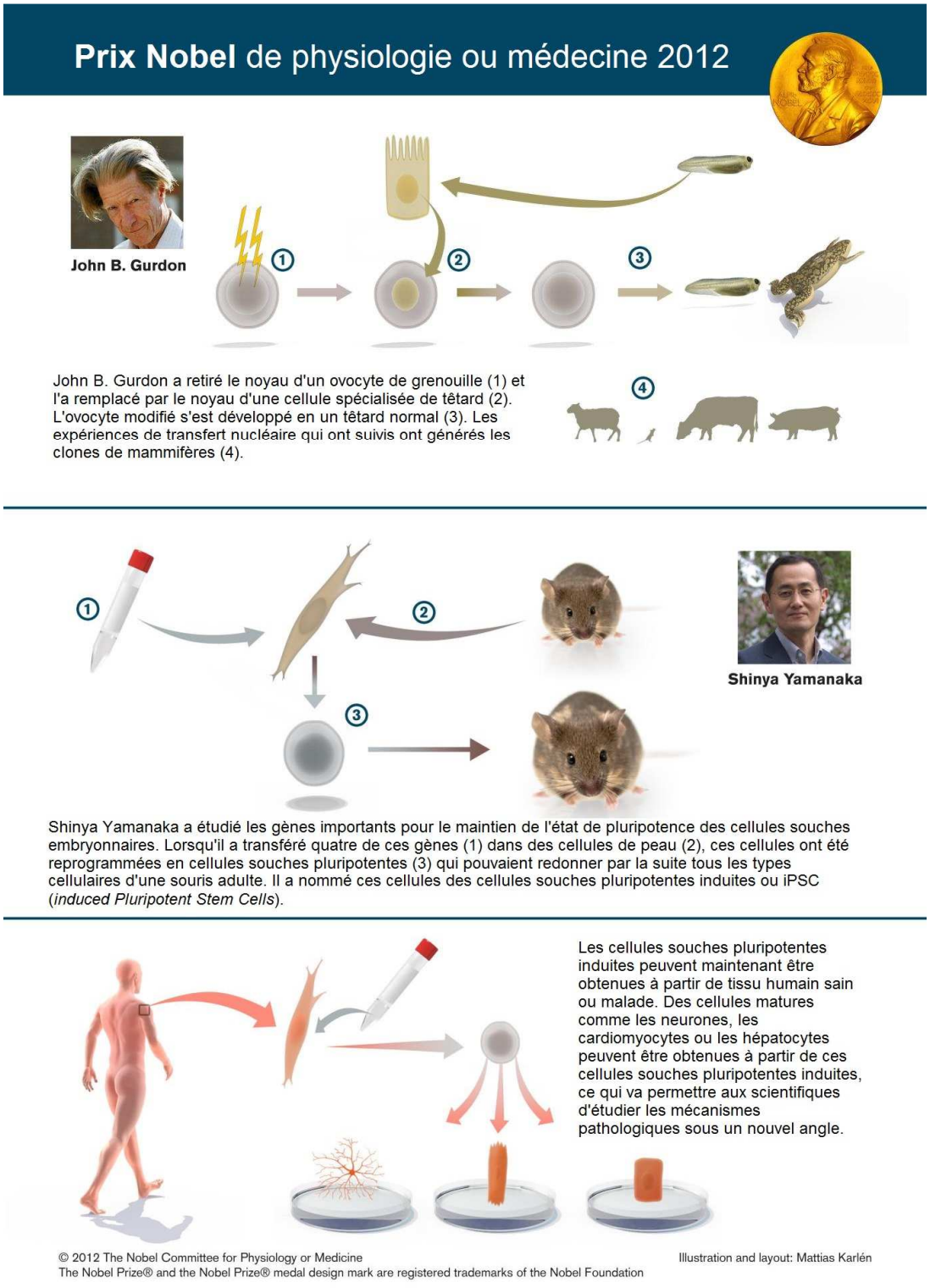
John Gurdon a émis l'hypothèse que le noyau de toute cellule contenait encore toute l'information nécessaire pour conduire le développement d'un organisme entier. Dans une expérience classique (Gurdon, 1962), il a remplacé le noyau d'un ovocyte de grenouille par le noyau d'une cellule spécialisée intestinale mature. Cet ovocyte modifié s'est développé en un têtard normal. Le noyau de la cellule mature avait donc encore toutes les informations nécessaires pour développer toutes les cellules de la grenouille, autrement dit le noyau d'une cellule spécialisée mature peut retourner à un état pluripotent immature. La découverte de Gurdon a d'abord été accueillie avec scepticisme, mais est devenu acceptée lorsqu'elle a été confirmée par d'autres scientifiques. Cela a lancé une recherche intense et la technique a été développée, conduisant finalement au clonage des mammifères.

Le travail de Shinya Yamanaka a porté sur les cellules souches embryonnaires, c'est à dire des cellules souches pluripotentes isolées à partir d'embryons et cultivées en laboratoire. L'équipe de Yamanaka a essayé de trouver les gènes qui maintiennent les cellules souches embryonnaires immatures. Lorsque plusieurs de ces gènes ont été identifiés, Yamanaka et ses collègues ont introduit ces gènes selon différentes combinaisons dans des cellules adultes, et ont regardé au microscope si l'un d'eux pouvait reprogrammer les cellules adultes en cellules souches pluripotentes. Ils ont finalement trouvé une combinaison qui a fonctionné : en introduisant quatre gènes codant des facteurs de transcription (Oct3/4, Sox2, c-Myc et Klf4), ils pouvaient reprogrammer leurs fibroblastes en cellules souches immatures ! Shinya Yamanaka a donc découvert plus de 40 ans après Gurdon (Takahashi et Yamanaka, 2006), que des cellules adultes matures intactes chez les souris pouvaient être reprogrammées pour devenir des cellules souches immatures pluripotentes, par l'addition de seulement quelques facteurs bien définis. Ces cellules, qui ont été appelées iPSCs (*induced Pluripotent Stem Cells*, cellules souches pluripotentes induites), pourraient à leur tour se développer en différents types de cellules matures telles que des fibroblastes, des cellules nerveuses et des cellules de l'intestin.

Ces découvertes révolutionnaires ont complètement changé notre vision du développement et de la spécialisation cellulaire. Nous savons maintenant que la cellule mature n'est pas confinée à jamais dans son état spécialisé. Des manuels ont été réécrits et de nouveaux champs de recherche ont été mis en place, surtout depuis que les cellules iPSCs peuvent également être préparées à partir

de cellules humaines (Takahashi *et al.*, 2007 ; Zaehres et Schöler, 2007). Ces cellules constituent de nouveaux outils précieux pour aider les scientifiques à comprendre les mécanismes de certaines maladies et ainsi offrir de nouvelles opportunités pour développer des méthodes de diagnostic et de thérapie (Figure 22).

Figure 22 : Illustration du prix Nobel de physiologie ou médecine de 2012

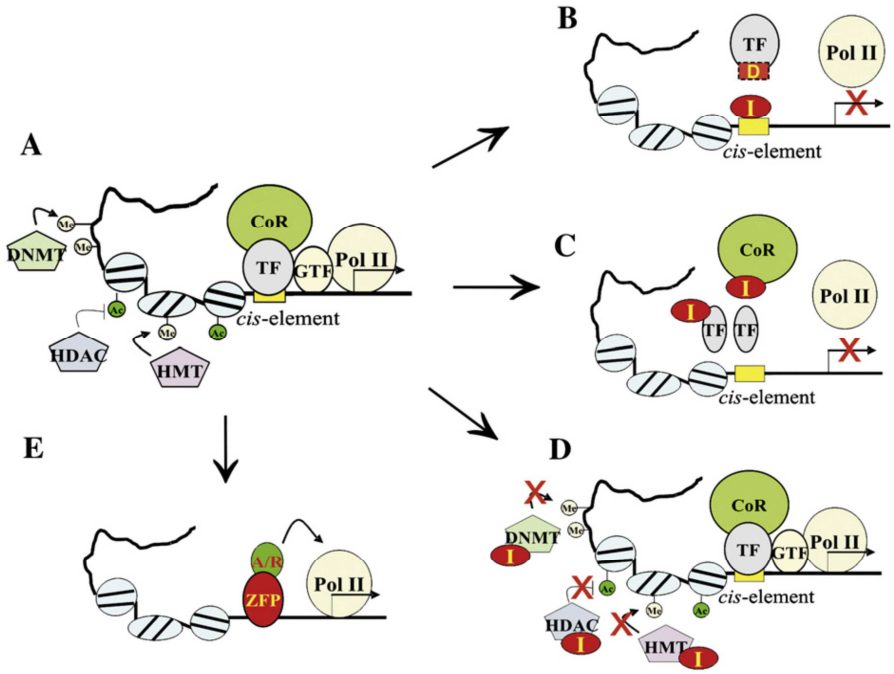


D'après ("The 2012 Nobel Prize in Physiology or Medicine - Press Release," 2012)

B. Les facteurs de transcription : nouvelles cibles thérapeutiques

Nous venons de voir que par leur importance biologique de nombreux facteurs de transcription étaient impliqués dans diverses maladies, notamment le cancer (défini comme une maladie génétique). Or de nombreuses anomalies transcriptionnelles ont été associées au cancer, dont certaines sont dues à des altérations de l'activité de facteurs de transcription, pouvant être dues ou non à des mutations (Yan et Higgins, 2013). Les facteurs de transcription sont donc considérés comme des cibles thérapeutiques très prometteuses, qui peuvent agir en stimulant la transcription de gènes spécifiques d'un effet bénéfique désiré, ou en inhibant la transcription de gènes impliqués dans une maladie. C'est la « thérapie transcriptionnelle », c'est-à-dire qui agit directement sur la transcription (Yan et Higgins, 2013). Certains médicaments sont utilisés en pratique, mais beaucoup sont encore en phase d'essais cliniques. En effet, les étapes de mises au point (essais *in vitro*, dans des modèles cellulaires, *in vivo*,…) sont longues du fait des actions diverses et variées et des mécanismes d'action complexes non encore maîtrisés des facteurs de transcription. Les différentes stratégies de la « thérapie transcriptionnelle » et donc les différents types de médicaments utilisés sont présentés en Figure 23.

Figure 23 : Stratégies thérapeutiques visant la régulation de la transcription



D'après (Yan et Higgins, 2013). **A :** La transcription est régulée à différents niveaux. **B :** Les petites molécules ou les polyamides (I) sont des compétiteurs des facteurs de transcription pour la fixation sur les séquences *cis*-régulatrices (exemple : Mithramycine pour c-Myc), tandis que les leurres (D) se lient à des facteurs de transcription et les empêchent de se lier à des promoteurs (exemple : des oligodésoxynucléotides double-brin qui miment les séquences de fixation de facteurs de transcription). **C :** Des peptides mimétiques ou de petites molécules (I) perturbent la dimérisation des facteurs de transcription (exemple : IA6B17 pour l'hétérodimérisation de c-Myc avec son partenaire Max) ou les interactions entre les facteurs de transcription et leurs co-régulateurs (exemple : KG-501 pour l'interaction CREB-p300). **D :** Les inhibiteurs des enzymes de modification de l'ADN ou des histones (I) modifient le paysage épigénétique, ce qui a des effets sur l'expression des gènes (exemple : l'azacytidine inhibe les DNMT). **E :** Des facteurs de transcription artificiels (ZFP) fusionnés avec des domaines d'activation ou de répression de la transcription (A/R) se lient aux promoteurs et modulent l'expression de gènes-cibles (voir plus loin).
TF : *transcription factor* (facteur de transcription) ; GTF : *General Transcription Factor* (facteur général de transcription) ; Pol II : l'ARN polymérase II ; CoR : co-régulateur ; DNMT : l'ADN méthyltransférase ; HDAC : histone désacétylase ; HMT : histone méthyltransférase ; I : agents de la transcription ciblés ; D : leurre d'un facteur de transcription ; Me : un groupement méthyle ; Ac : un groupement acétyle ; ZFP : *zinc-finger protein* (protéine en doigt de zinc) ; A/R : domaine d'activation ou de répression transcriptionnelle.

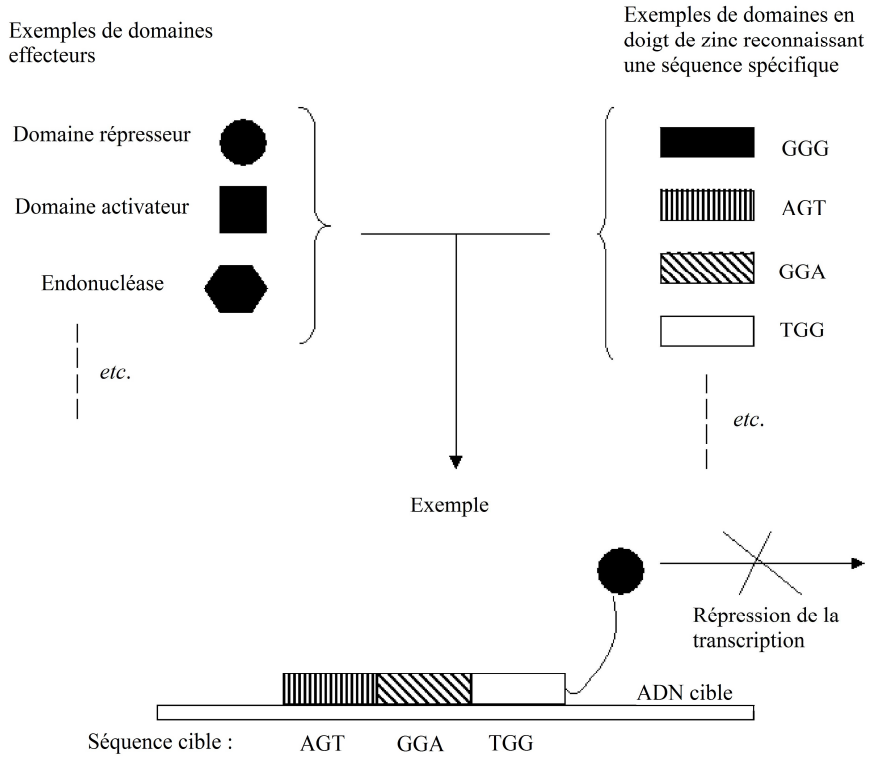
Les thérapeutiques les plus développées concernent à ce jour les récepteurs nucléaires dont on connaît le ligand (Dunker et Uversky, 2010) : par exemple les stéroïdes anabolisants ou la dexaméthasone pour le récepteur des glucocorticoïdes (ciblée dans les maladies inflammatoires), les thiazolidinediones pour le PPARgamma (*Peroxysome Proliferator-activated Receptor gamma*, ciblé dans le diabète de type II), le tamoxifène pour le récepteur aux œstrogènes (ciblé dans le cancer du sein), et le bicalutamide pour le récepteurs aux androgènes (ciblé dans le cancer de la prostate). Parfois, certains médicaments modulent les fonctions des facteurs de transcription indirectement, par l'intermédiaire de nombreuses cascades de signalisation (comme les inhibiteurs des voies de signalisation activant le NF-κB par exemple).

De petites molécules sont maintenant construites pour reconnaître et bloquer spécifiquement les sites de liaison à l'ADN ou à un ligand. Des recherches sont en cours sur les récepteurs nucléaires (Dunker et Uversky, 2010).

Certaines recherches, plus complexes, sont aussi axées sur des régions particulières de certaines protéines : les IDR (*Intrinsically Disordered Regions*), des régions structurellement instables, flexibles, ou sujettes à des changements de conformation corrélées à leur fonction (qui vont participer à des interactions protéines-protéines avec des cofacteurs par exemple). Les premiers résultats sur c-Myc ouvrent un nouveau champ de recherche : il serait possible, grâce à de petites molécules (comme IA6B17), d'empêcher l'hétérodimérisation de c-Myc (*v-myc avian myelocytomatosis viral oncogene homolog*, facteur de transcription surexprimé dans de nombreux cancers humains) avec son partenaire Max (*MYC associated factor X*, facteur de transcription possédant un motif de type fermeture éclair à leucines) dans le but d'inactiver c-Myc (Metallo, 2010).

Un grand intérêt est porté depuis quelques années sur la construction de facteurs de transcription artificiels (Gommans *et al.*, 2005). Ces protéines chimériques sont construites sur la base des protéines à doigt de zinc (Figure 24) : elles contiennent plusieurs motifs en doigt de zinc de type Cys_2/His_2 en tandem. Chaque motif contient environ 30 acides aminés mais seulement quelques-uns sont nécessaires pour lier trois ou quatre bases successives au niveau du grand sillon de l'ADN. Une modification d'acide aminé au niveau du site de reconnaissance d'un doigt de zinc change le motif de liaison à l'ADN. À ce jour, plusieurs motifs à doigt de zinc reconnaissant trois paires de bases ont été caractérisés. Il est possible d'assembler plusieurs domaines en doigt de zinc en tandem pour générer un facteur de transcription artificiel pouvant se lier à une séquence d'ADN bien spécifique. Par exemple, une protéine à doigt de zinc générée en reliant six doigts de zinc reconnaît un site de 18 paires de bases, ce qui constitue théoriquement un site unique et spécifique dans le génome humain. On ajoute à ces protéines en doigt de zinc un domaine d'activation de la transcription (par exemple, le domaine de transactivation de la protéine du virus de l'herpès simplex VP16, *virus protein 16*, ou celui de la sous-unité p65 du facteur de transcription humain NF-κB) qui permet l'activation de l'expression spécifique du(des) gène(s)-cible(s). On peut également ajouter un domaine de répression de la transcription (par exemple le domaine KRAB, *Krüppel associated box domain*, domaine présent dans approximativement 400 facteurs de transcription en doigt de zinc chez l'homme). De grandes avancées ont été faites ces dernières années et ces protéines deviennent de plus en plus spécifiques. Par exemple, un facteur de transcription artificiel peut être construit pour se lier spécifiquement au site de liaison à l'ADN de la protéine p53, site que l'on retrouve dans le promoteur du gène *Bax* : le facteur de transcription artificiel peut alors activer l'expression de *Bax*, *BCL2(B-cell chronic lymphocytic leukemia/lymphoma 2)-associated X protein*, qui intervient dans la voie de l'apoptose (Yan et Higgins, 2013). Remarque : les motifs de liaison à l'ADN en doigt de zinc sont également utilisés pour construire des enzymes de restriction spécifiques, les *zinc-finger nucleases*, qui permettent entre autre de faire de la mutagenèse dirigée.

Figure 24 : Construction d'un facteur de transcription artificiel en doigt de zinc



D'après (Gommans *et al.*, 2005). Il est possible de coupler une grande variété de domaines effecteurs à des domaines de liaison à l'ADN (en effet ces deux domaines étaient interchangeables, c'est une propriété des facteurs de transcription). De plus les domaines en doigt de zinc reconnaissant une séquence spécifique sont aussi modulables et peuvent être assemblés en tandem pour reconnaître une séquence plus étendue.

Les applications de ces molécules régulant directement la transcription commencent à voir le jour, mais on est encore loin de la thérapie génique. Les problèmes majeurs rencontrés sont la disponibilité de la molécule pour les cellules ciblées (la molécule doit atteindre la tumeur par exemple), la stabilité de la molécule et la spécificité de la molécule (Yan et Higgins, 2013). Le but ultime est de cibler un seul ou quelques gènes dans un tissu choisi, sans altérer ni les autres gènes ni les autres tissus, ce qui n'est pas simple avec les facteurs de transcription. On peut supposer que les applications les plus directes se feront via les cellules souches pluripotentes induites.

IV. Méthodes d'analyse spécifiques pour l'étude des facteurs de transcription

Comme nous l'avons vu précédemment, les mécanismes de régulation de l'expression des gènes eucaryotes par les facteurs de transcription ont comme point de départ l'interaction de protéines régulatrices avec des séquences d'ADN spécifiques. Au cours de l'étude de la régulation d'un gène, il est donc particulièrement important d'une part de pouvoir mettre en évidence ces interactions, et d'autre part de pouvoir démontrer la spécificité et le rôle régulateur de l'interaction.

Je m'attarderai surtout sur les méthodes spécifiques. Les techniques plus générales (qui ne visent souvent pas que l'étude des facteurs de transcription) seront citées mais non développées, tout comme les techniques classiques de biologie moléculaire (PCR, transfection, production de protéines,…).

Plusieurs expériences ayant montré que les facteurs de transcription peuvent se lier étroitement à l'ADN tant à l'intérieur des cellules que *in vitro* (Phillips et Hoopes, 2008), il existe maintenant des techniques *in vitro* et *in vivo* pour caractériser les facteurs de transcription, mais l'étape de validation finale de l'effet devra toujours être réalisée *in vivo* (ou au moins dans un modèle cellulaire).

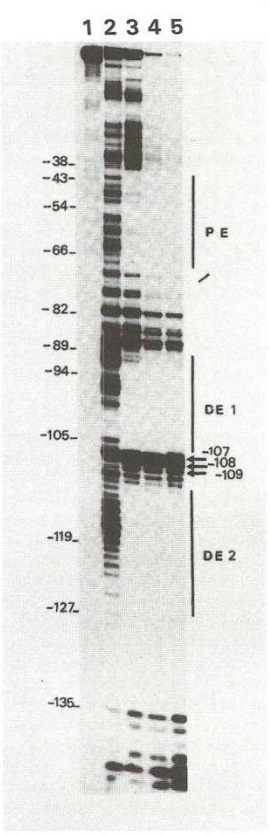
A. Mise en évidence d'interactions ADN/protéine (régulation en *trans*)

1. L'empreinte à la DNAse I (*footprinting*)

D'après (Kaplan et Delpech, 2007).

Cette technique permet de déterminer le siège de l'interaction entre facteurs protéiques et ADN. Elle repose sur le fait que l'interaction d'une protéine avec une séquence d'ADN protège cette séquence vis-à-vis des DNAses. Dans la pratique le fragment d'ADN exploré, dont la longueur ne doit pas dépasser quelques centaines de paires de bases, est préalablement marqué avec du ^{32}P à l'une de ses extrémités 5' ou 3'. Il est incubé en présence d'un extrait de protéines où sont censées se trouver les protéines qui interagissent (dans la pratique il s'agit en général d'un extrait des protéines nucléaires). Puis on effectue une digestion par la DNAse I. En parallèle, une incubation sans extrait protéique est effectuée, elle servira de témoin. Les produits de digestion sont enfin analysés par électrophorèse sur un grand gel de polyacrylamide, et une autoradiographie est pratiquée (Figure 25). La fraction d'ADN non digéré que l'on visualise sur l'autoradiographie correspond aux séquences d'ADN interagissant avec des protéines, donc à l'empreinte de la protéine, d'où l'appellation d'empreinte à la DNAse I ou *footprinting*. Cette technique devient de plus en plus obsolète car imprécise aujourd'hui, et peu efficace dans la pratique.

Figure 25 : Résultat d'une expérience d'empreinte à la DNAse I



D'après (Kaplan et Delpech, 2007), cliché de M. Raymondjean, Institut Cochin, Paris.

1 : ADN témoin sans DNase.
2 : ADN témoin incubé sans protéines. Dans ce canal, on observe une série continue de bandes rapprochées, chacune correspondant à un fragment d'ADN restant après la coupure par la DNAse. La DNase coupant peu et au hasard, sur l'ensemble des fragments le nombre de coupures après chacune des bases est statistiquement identique.
3, 4 et 5 : ADN incubé avec une concentration croissante de DNase I. Dans ce canal, une ou plusieurs zones plus ou moins longues sont dépourvues de bandes (zones PE, DE1 et DE2). Cela indique que la séquence d'ADN correspondante n'a pas été attaquée par la DNase, donc était protégée par des protéines fixées. Sur les bords des zones protégées il est fréquent d'observer une ou deux bandes plus intenses indiquant que les coupures ont été plus fréquentes en cet endroit, elles correspondent à des zones hypersensibles, l'accès de la DNase à l'ADN étant facilité soit par une structure particulière de l'ADN, soit par une interaction avec la protéine fixée à l'ADN. La zone protégée peut être déterminée à la base près. Pour faciliter le repérage au sein de la séquence, une détermination de séquence est effectuée sur le même gel de part et d'autre.

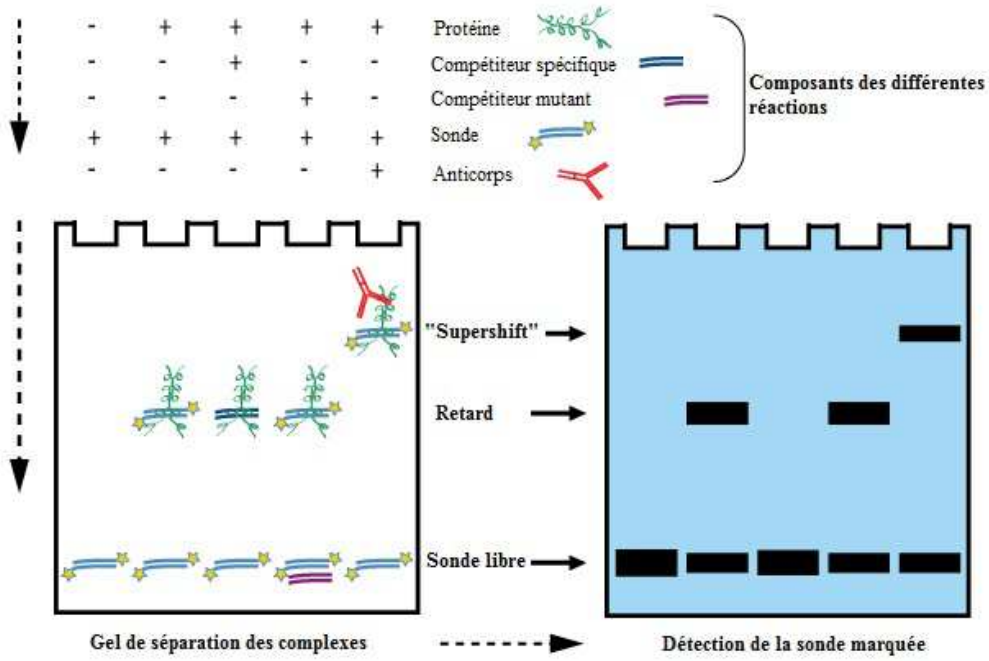
2. <u>Le retard de migration sur gel</u>

D'après (Kaplan et Delpech, 2007).

Cette technique permet d'objectiver la présence et la spécificité d'une interaction physique entre une protéine et une séquence définie d'ADN. À l'origine, on marquait au ^{32}P soit un oligonucléotide, soit la séquence d'ADN soupçonnée de posséder un site d'interaction avec une protéine. Le fragment d'ADN radioactif est incubé en présence d'un extrait protéique contenant le facteur qui est supposé interagir. La source de cette protéine peut même être un extrait nucléaire brut, aucune pureté particulière n'étant requise pour ce genre d'expérience. Après incubation, l'échantillon est ensuite déposé sur un gel de polyacrylamide non dénaturant à mailles assez lâches (du fait de la taille des complexes à séparer). L'interaction de la protéine avec la séquence d'ADN va augmenter sa masse moléculaire apparente, donc retarder sa migration dans le gel. Les protéines régulatrices étant toujours présentes en très faible concentration, seule une faible fraction de la séquence sera retardée. Le retardement de cette fraction s'objectivera par la présence d'une bande correspondant à un matériel de masse moléculaire plus élevée. L'ADN étant par nature une molécule très chargée, la plupart des interactions sont de type non spécifique. Afin de les éliminer, une série d'incubations sont effectuées en présence de concentrations croissantes d'un compétiteur non spécifique, en général du poly(dI-dC). L'interaction sera considérée comme spécifique si l'intensité de la bande retardée n'est pas modifiée par la présence du compétiteur aspécifique. En cas de non spécificité une décroissance progressive de la bande avec les concentrations croissantes de compétiteur est observée. Un deuxième contrôle consiste à utiliser la séquence étudiée comme compétiteur. Si le phénomène est spécifique, le signal doit disparaître au fur et à mesure de l'augmentation du compétiteur (Figure 26).

Afin d'éviter l'utilisation de la radioactivité, il est possible aujourd'hui de remplacer le ^{32}P par d'autres marqueurs avec des performances analogues: marqueur fluorescent (Fuji), biotine,…

Figure 26 : Principe de l'expérience de retard sur gel



D'après le site de Thermo Scientific (http://www.thermoscientific.com/). Le retard sur gel consiste en trois étapes clé : (1) réactions de liaison, (2) électrophorèses, (3) détection de la sonde. L'ordre dans lequel on ajoute les composants pour la réaction de liaison est souvent critique. Cet exemple idéal montre une complète disparition du complexe protéine-sonde avec l'addition d'un compétiteur spécifique, et une bande plus haute lors de l'addition d'un anticorps dirigé spécifiquement contre la protéine (cette réaction s'appelle un « supershift » et montre la spécificité de la protéine vis-à-vis de la séquence). Cependant, on voit plus souvent une réduction d'intensité qu'une complète disparition de la bande.

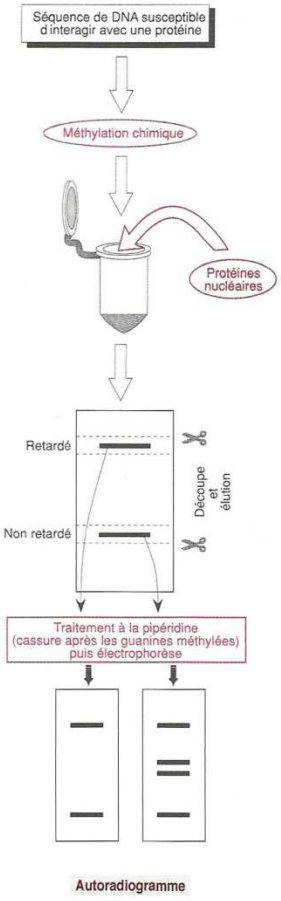
3. L'interférence de méthylation

D'après (Kaplan et Delpech, 2007).

Cette technique, qui vient en complément des deux précédentes, permet d'affiner l'étude de l'interaction entre une protéine et une séquence d'ADN. Elle permet de déterminer quelles sont les guanines de la séquence nécessaires à son interaction avec la protéine. En pratique, le fragment d'ADN susceptible d'interagir avec une protéine est marqué au ^{32}P à l'une de ses extrémités, puis est partiellement méthylé sur ses guanines avec du diméthylsulfate. La réaction est conduite de telle façon que, au hasard, seules quelques guanines sont méthylées. Si la méthylation s'est effectuée sur une guanine impliquée dans une interaction avec une protéine cette interaction ne peut plus se contracter. Il est alors pratiqué une expérience de retard sur gel comme décrit dans le paragraphe précédent. Les bandes retardée et non retardée sont découpées et l'ADN est élué. Une détermination de séquence limitée aux seules guanines est effectuée par la technique de Maxam et Gilbert (technique de séquençage). Les séquences obtenues sont comparées. Comme la méthylation d'une guanine empêche son interaction avec une protéine, les fragments méthylés sur une guanine impliquée dans l'interaction ne seront plus retardés dans le gel. Il en résulte que dans l'analyse de la séquence, la bande correspondant à cette guanine sera absente. La comparaison de la séquence des

G entre fragments retardé et non retardé permet donc de localiser les guanines impliquées (Figure 27).

Figure 27 : La technique d'interférence de méthylation



D'après (Kaplan et Delpech, 2007)
Lorsqu'une guanine interagissant spécifiquement avec une protéine est méthylée, l'interaction avec cette protéine n'est plus possible. Dans une expérience de retard sur gel, l'ADN qui la contient n'est plus retardé. Pour mettre ce phénomène en évidence l'ADN est méthylé par le disulfate de méthyl dans des conditions douces de telle sorte qu'une seule des guanines de la séquence soit méthylée. Il est ensuite pratiqué une expérience de retard sur gel. Les bandes retardées et non retardées sont traitées à la pipéridine qui coupe après les guanines méthylées (principe de la technique de Maxam et Gilbert). Les fragments sont analysés par électrophorèse et autoradiographie. Les bandes correspondant aux sites d'interaction ne sont pas observées dans le fragment retardé (si elles avaient été méthylées, le fragment n'aurait pas été retardé).

4. L'empreinte à la DNAse I (*footprinting*) *in vivo*

D'après (Kaplan et Delpech, 2007).

Il est impossible d'exclure que les interactions entre une protéine et une séquence d'ADN cible, observées *in vitro*, ne résultent pas d'artefacts. La technique d'empreinte à la DNAse I (*footprinting*) *in vivo* permet de démontrer la réalité des interactions observées *in vitro*.

Des cellules en culture dans lesquelles le gène, dont on souhaite étudier la régulation, est exprimé sont trypsinisées puis traitées, dans des conditions très douces, par le diméthyl sulfate. Ce composé diffuse dans le noyau et méthyle l'azote 7 des guanines qui ne sont pas protégées par une interaction avec une protéine. Compte tenu des conditions expérimentales, seules quelques-unes des guanines accessibles sont méthylées. Les noyaux sont purifiés et l'ADN est extrait. L'ADN est ensuite digéré par une enzyme de restriction puis traité à la pipéridine, composé qui casse l'ADN au niveau des guanines méthylées. Les régions de l'ADN qui étaient protégées par des protéines et qui n'ont pas été méthylées ne seront jamais détruites, alors que celles qui n'étaient pas protégées ont été partiellement méthylées et seront attaquées par la pipéridine. Suivant le même principe statistique que celui utilisé dans la technique de Maxam et Gilbert, il sera obtenu autant de types de fragments que l'ADN contient de guanines méthylées. Une électrophorèse en gel de polyacrylamide

est ensuite pratiquée afin de séparer les fragments en fonction de leur taille, puis il est effectué un transfert sur une membrane de nylon. Cette membrane est préhybridée puis hybridée avec une sonde marquée au ^{32}P correspondant à la région du gène où l'on pense qu'une protéine régulatrice pourrait se fixer. Le profil observé à l'autoradiographie est proche de celui obtenu dans les expériences d'interférence de méthylation. De l'ADN témoin est traité dans les mêmes conditions. Il permet de repérer les guanines de la séquence et ainsi de différencier celles qui n'étaient pas protégées (bande présente à la fois dans l'échantillon et dans le témoin) de celles qui étaient protégées (bande présente dans le témoin et absente dans l'échantillon) et donc interagissaient avec la protéine.

5. Pontage aux ultraviolets entre une séquence d'ADN et une protéine

D'après (Kaplan et Delpech, 2007).

Le but de cette technique est de déterminer la masse moléculaire d'une protéine se fixant à l'ADN sur une séquence cible, ce qui ne pouvait être fait par la technique de retard sur gel (car cette dernière est réalisée en conditions non-dénaturantes). Le principe consiste à lier de manière covalente la protéine à sa séquence cible, puis de déterminer la masse moléculaire du complexe par électrophorèse SDS-PAGE (*SDS-PolyAcrylamide Gel Electrophoresis*). Dans la pratique, la première étape consiste à synthétiser la séquence cible en utilisant de la 5-bromodésoxyuridine (BrdU) à la place de la thymine. Le premier brin de la séquence cible est tout d'abord synthétisé de manière classique par un synthétiseur d'oligonucléotides. Cet oligonucléotide doit aussi comprendre les séquences adjacentes à celle étudiée, sa longueur doit être d'au moins une cinquantaine de paires de bases. Un second oligonucléotide, plus court (une dizaine de paires de bases), dont la séquence est complémentaire de l'extrémité 3' du premier oligonucléotide, est synthétisé puis hybridé. Il sert d'amorce pour l'ADN polymérase I qui synthétise le reste du brin. Dans le milieu d'incubation le dTTP est remplacé par de la 5-bromodésoxyuridine triphosphate et le dATP est remplacé par du $^{alpha32}P$ dATP. L'ADN double brin ainsi synthétisé est radioactif, ce qui permettra de le repérer. Il contient du BrdU qui permettra de contracter des liaisons covalentes avec les protéines qui lui sont associées, et ce par simple irradiation aux UV. Cet ADN ainsi synthétisé et marqué est utilisé pour réaliser une expérience de retard sur gel. Après migration, le gel est irradié par des UV à 302 nm, puis il est pratiqué une autoradiographie afin de repérer les bandes radioactives, lesquelles sont découpées. Les morceaux d'acrylamide sont incubés dans un tampon dénaturant, les protéines ainsi éluées du gel sont déposées sur un gel de polyacrylamide SDS. La masse moléculaire des complexes ADN/protéines est évaluée, et celle de la protéine est extrapolée en déduisant la masse moléculaire du fragment d'ADN.

B. La caractérisation des séquences possédant un rôle régulateur (régulation en *cis*)

1. Par précipitation de chromatine : ChIP (*Chromatin ImmunoPrecipitation*) et sa variante le ChIP-on-chip

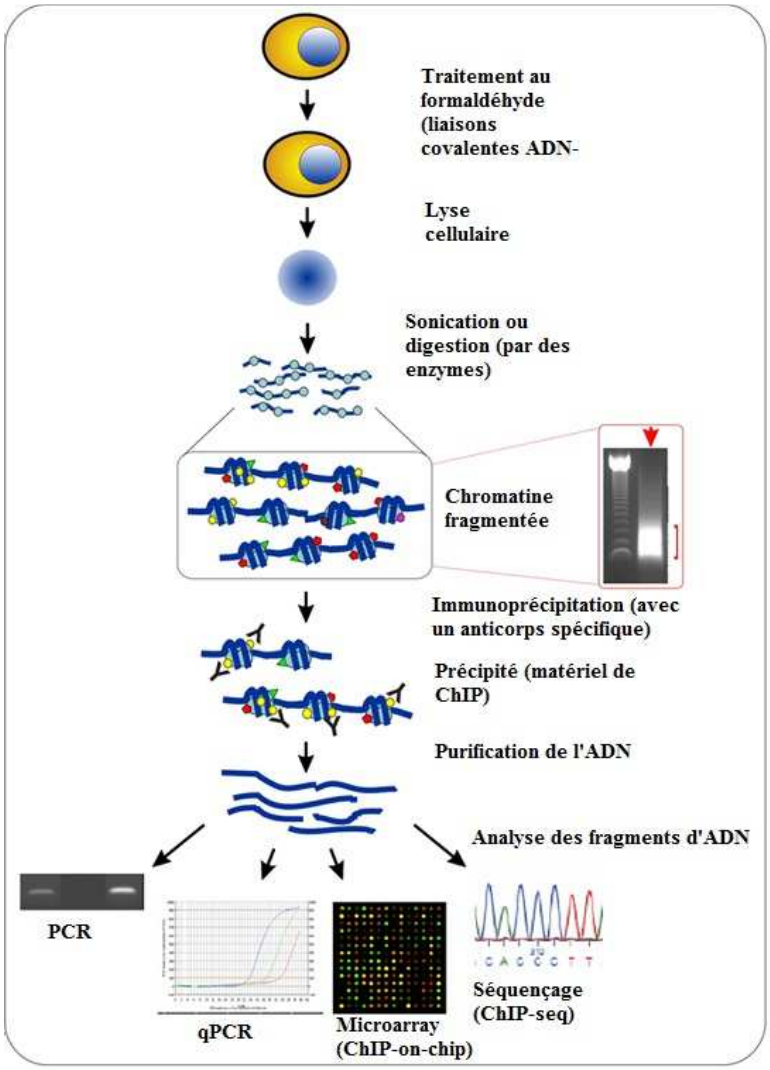
D'après (Kaplan et Delpech, 2007).

Il est possible de déterminer la séquence d'ADN sur laquelle un facteur de transcription se fixe par la méthode de précipitation de la chromatine, dite ChIP. Pour cela des cellules en culture ou issues d'un tissus sont perméabilisées puis traitées avec du formaldéhyde, un agent qui crée des liaisons covalentes entre les groupements aminés des protéines (lysines, arginines) ainsi qu'entre les groupements aminés des protéines et les groupements aminés des bases de l'ADN. Il faut pour cela que les groupements aminés soient suffisamment proches les uns des autres et qu'il existe donc une (des) interaction(s) protéine/protéine ou protéine/ADN. Le caractère covalent de la liaison rend les

complexes formés particulièrement stables, ils ne seront donc pas détruits par les traitements ultérieurs utilisés pour les purifier. La chromatine est ensuite extraite, puis traitée par les ultrasons (sonication), afin de casser l'ADN en petits fragments. Un anticorps dirigé contre l'une des protéines interagissant avec l'ADN est ensuite utilisé afin d'immunoprécipiter les complexes. Une fois les complexes récupérés, ils sont traités de manière à détruire les liaisons covalentes qui avaient été créées par le formaldéhyde. Le type de traitement est fonction des études que l'on veut réaliser ensuite. Si l'on souhaite analyser à la fois l'ADN et les protéines qui constituent le complexe, il faut réaliser un clivage doux qui n'altère pas les protéines. Si l'on ne s'intéresse qu'à l'ADN (ce qui est souvent le cas), les protéines sont détruites avec de la protéinase K, une protéase particulièrement efficace, et l'ADN est extrait. Dans les premières utilisations de la technique l'analyse de l'ADN récupéré se faisait par Southern-blot, clonage ou PCR. L'approche était lourde et nécessitait de connaître par avance, au moins partiellement, les séquences recherchées. La technique ChIP a été ensuite améliorée : les fragments d'ADN récupérés (d'une longueur de 200 à 1 000 pb) peuvent être clonés et séquencés pour rechercher un motif particulier (ChIP-seq), ou ils peuvent être analysés avec des puces à ADN (technique du ChIP-on-chip, voir Figure 28). L'ADN récupéré est marqué avec un fluorochrome et hybridé sur une puce à ADN contenant des séquences potentielles d'intérêt (puces de promoteurs par exemple). Le repérage des spots de la puce où une hybridation s'est produite permet d'identifier les séquences d'ADN qui étaient en interaction avec le complexe protéique dans la cellule. Les études réalisées sont le plus souvent des études comparatives entre des cellules qui ont subi des stimulations différentes ou qui sont dans des états de différenciation différents. Dans de tels cas, afin de simplifier les expérimentations, les ADN provenant des différents types cellulaires peuvent être marqués avec des fluorochromes de différentes couleurs.

Cette technique, simple dans son principe, est de réalisation délicate. Le temps de traitement par le formaldéhyde est critique : si le temps de traitement est trop long (les meilleurs résultats sont obtenus pour des temps de traitement compris entre quelques minutes et trente minutes), les complexes sont tellement stables qu'ils ne peuvent plus être détruits, et il y a un risque que les anticorps ne reconnaissent plus la protéine contre laquelle ils sont dirigés. Il ne faut pas espérer caractériser d'emblée la liste des séquences du génome qui sont la cible d'un facteur protéique donné. Le bruit de fond, et donc le risque de faux positifs, est souvent élevé car de nombreux contaminant sont entraînés lors de l'immuno-purification. Les variations non significatives sont nombreuses. Il convient donc de multiplier les expérimentations et de ne prendre en compte que les résultats significativement reproductibles. L'aide d'une équipe de bio-informaticiens est hautement souhaitable pour analyser et interpréter les résultats.

Figure 28 : Les différentes techniques de ChIP



D'après (Collas et Dahl, 2008)
La ChIP est une technique d'identification des séquences d'ADN spécifiques qui sont liés, *in vivo*, à des protéines d'intérêt. Elle implique la fixation, par du formaldéhyde, de la chromatine à des protéines liant l'ADN (en général des facteurs de transcription). Après sonication de l'ADN en petits fragments, les complexes spécifiques protéines-ADN sont isolés par immunoprécipitation grâce à des anticorps spécifiques de la protéine d'intérêt. Puis l'ADN isolé à partir du complexe peut être :

- Amplifié par PCR (ce qui suppose d'avoir des amorces) ;
- Identifié par séquençage (ChIP-seq) ;
- Quantifié par qPCR ;
- Hybridé sur une puce à ADN (ChIP-on-chip).

2. PCR-sélection

Cette méthode a été décrite en 2008 par une équipe qui recherchait une séquence consensus de fixation pour le facteur de transcription FOXL2 (Benayoun *et al.*, 2008). La description de cette méthode sera développée dans la partie expérimentale de cette thèse (voir II.D). La méthode est assez similaire à une ChIP : le principe est de purifier, à l'aide d'un anticorps, des oligonucléotides sur lesquels un facteur transcriptionnel se fixe, au sein d'une banque d'oligonucléotides aléatoires. S'en suit une analyse bioinformatique pour identifier un motif commun au sein des oligonucléotides purifiés. Comparé à la ChIP, cette technique est moins cher, plus simple à mettre en œuvre et plus simple d'analyse ; en effet, les oligonucléotides aléatoires utilisés et purifiés font tous 76 pb, alors que les fragments récupérés à la suite d'une ChIP font entre 200 et 1 000 pb et sont donc plus difficile à aligner pour identifier un motif, même avec un logiciel (voir plus haut). La PCR-sélection a cependant l'inconvénient de se faire *in vitro* et non *in vivo* comme dans la ChIP.

C. Méthodes d'analyse protéiques

D'après (Kaplan et Delpech, 2007).

Les facteurs de transcription sont des protéines, et en cela leur analyse nous fait entrer dans le vaste monde de la protéomique (que je ne vais pas détailler dans cette thèse). Les méthodes d'étude des protéines sont très nombreuses, et je ne décrirai ici que celles qui sont utilisées pour analyser l'expression d'un facteur de transcription au niveau protéique. Il est également possible de caractériser les partenaires d'une interaction protéique, en utilisant les outils de la protéomique.

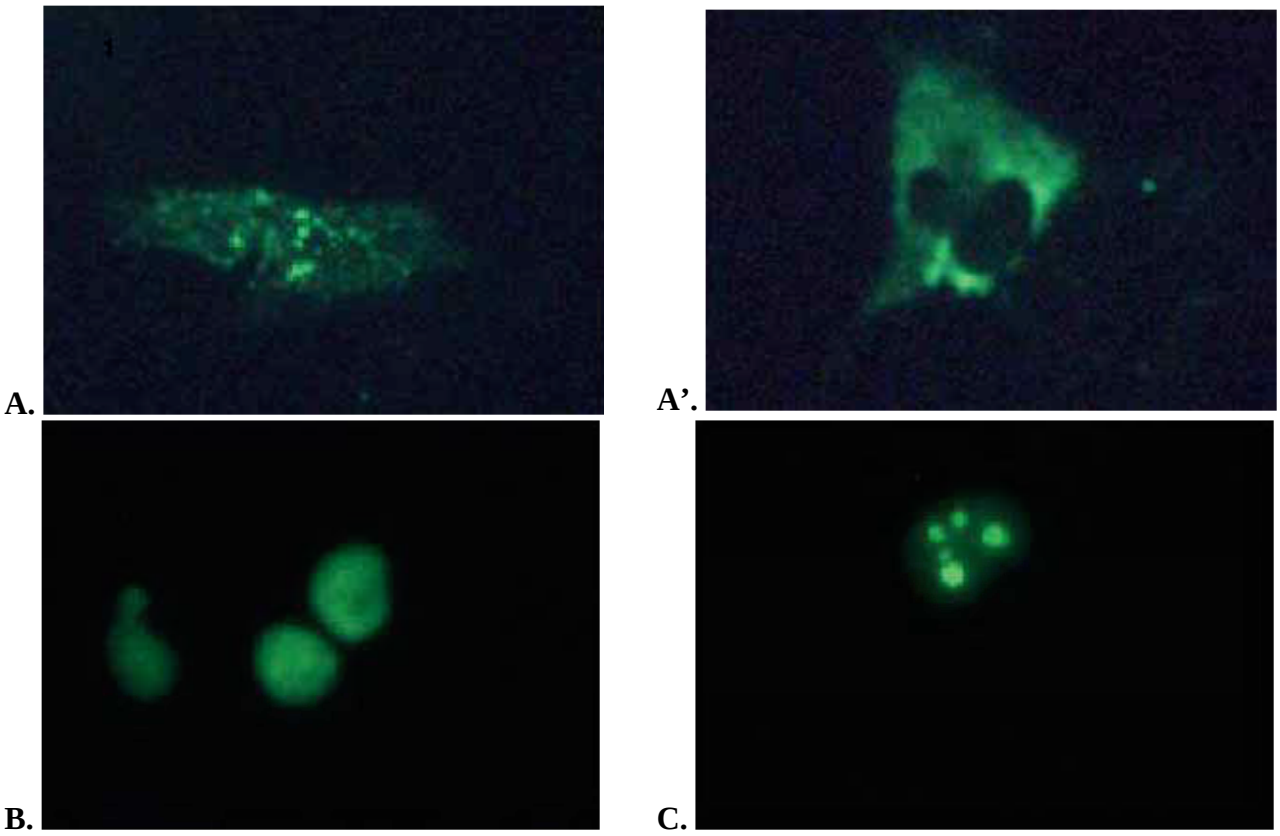
1. La localisation et le suivi d'une protéine au sein de la cellule

Lorsqu'un nouveau facteur de transcription a été caractérisé, le décryptage des mécanismes physiologiques et/ou physiopathologiques de son action reste à déterminer. Dans un premier temps il est plus raisonnable d'utiliser des approches simples qui consistent, par exemple, à repérer la protéine et la suivre au sein de la cellule lorsque cette dernière subit un stimulus. Si l'on dispose d'un anticorps, l'approche la plus simple est immunocytologique. Si l'on n'en possède pas, il est possible de la repérer et la suivre en la couplant à une protéine naturellement fluorescente ou un peptide pour lequel un anticorps existe (tag-His, tag-Flag, …).

Si l'approche expérimentale utilisée fait que l'on ne dispose pas d'anticorps dirigés contre la protéine que l'on souhaite étudier, il est possible d'en produire à partir des données sur la séquence de son ADNc (des sociétés spécialisées proposent de réaliser ce travail pour un coût raisonnable). Des logiciels ont été développés afin de repérer dans la séquence d'une protéine les régions qui, d'un point de vue théorique, devraient être le plus immunogènes. Même si ces logiciels sont maintenant optimisés, les résultats obtenus ne sont pas toujours à la hauteur des espérances et pour certaines protéines les anticorps sont peu spécifiques, et de ce fait parfois inutilisables. En cas de succès l'anticorps est marqué avec un fluorochrome ou plus rarement une enzyme afin de pouvoir repérer dans la cellule les complexes antigène/anticorps formés.

S'il n'est pas possible de disposer d'anticorps une alternative pour localiser et suivre une protéine au sein d'une cellule consiste à associer l'ADNc qui la code à celui d'une protéine naturellement fluorescente (ou un Flag), et introduire l'ADNc hybride dans un vecteur contenant toutes les séquences nécessaires pour une forte expression. Le vecteur recombinant est ensuite transfecté dans des cellules en culture. Il est naturellement important que le type cellulaire utilisé soit représentatif de la cellule dans laquelle on souhaite étudier la localisation ou le devenir de la protéine après sa synthèse, ou encore à la suite d'une stimulation. Ce type d'approche est aussi utilisable avec les animaux transgéniques où il est ainsi possible d'analyser l'expression d'une ou plusieurs protéines (GFP de différentes couleurs, voir plus loin) au cours du développement. La première protéine naturellement fluorescente a été découverte en 1961 dans l'algue *Aequorea victoria* : il s'agit de la *Green Fluorescent Protein* (GFP), une protéine constituée de 238 acides aminés. Le pic d'excitation est à 395 nm, c'est-à-dire dans le proche UV, et le pic d'émission est à 509 nm, c'est-à-dire dans le vert. Plusieurs mutations ont été introduites dans le gène de la GFP afin d'optimiser sa synthèse, sa stabilité ou de modifier ses longueurs d'onde d'excitation et/ou d'émission. La modification de la longueur d'onde d'émission permet d'obtenir des GFP de différentes couleurs : cyan, bleu, jaune. Il est ainsi possible d'analyser en même temps l'expression de plusieurs types de protéines au sein d'une même cellule, chaque couleur étant spécifique d'un type de protéine. Un inconvénient de la GFP est que sa fixation modifie la masse moléculaire de la protéine d'intérêt et peut induire une modification de sa conformation et de ses propriétés physico-chimiques. De ce fait il n'est jamais exclu que les propriétés biologiques de la protéine d'intérêt ne soient pas altérées. La validité des résultats obtenus doit donc être considérée avec une certaine réserve (Figure 29).

Figure 29 : Localisation des isoformes d'une protéine à l'aide de la GFP



D'après (van Dijk *et al.*, 2005). Analyse de la localisation cellulaire des différentes isoformes de STOX1 dans des cellules SGHPL5 (cytotrophoblastes extravilleux diploïdes normaux immortalisés par une transformation avec SV40, il s'agit d'un modèle cellulaire pour étudier les cytotrophoblastes extravilleux ayant la capacité de fusionner pour former des syncytia) par couplage à la GFP. L'expression nucléaire de l'isoforme A dans les cytotrophoblastes extravilleux est restreinte aux cellules polyploïdes (A), l'expression cytoplasmique, aux cellules diploïdes (A'). L'expression nucléaire et cytoplasmique est exclusive. Les isoformes B (B) et C (C) sont localisés dans les noyaux seulement. L'isoforme C est localisée dans le nucléole (nous verrons plus tard qu'il possède un domaine de localisation nucléaire ou NLS).

2. <u>Analyse d'interactions protéine-protéine</u>

Les mécanismes de la vie sont basés sur des cascades d'interactions. La caractérisation des partenaires de ces interactions représente donc un problème biologique dont l'enjeu est majeur. Les techniques permettant d'identifier les partenaires d'une protéine sont nombreuses et complexes, leur description détaillée sort du cadre de cette thèse.

Deux techniques permettant de cloner l'ADNc des partenaires d'une protéine donnée ont été initialement développées pour cela : la technique double hybride (système utilisant un facteur de transcription de levure, en général Gal4 de *Saccharomyces cerevisiae*, et basé sur le sur le fait que les deux domaines des facteurs transcriptionnels eucaryotes, le domaine d'interaction avec l'ADN et le domaine d'activation de la transcription, sont interchangeables) et la technique du phage display (système utilisant une banque de phages recombinants). Ces techniques sont efficaces mais très lourdes, elles conduisent souvent à des faux positifs et ne sont donc pas adaptées à la caractérisation des facteurs transcriptionnels.

Pour caractériser des interactions protéine/protéine, on préfère aujourd'hui utiliser les outils de la protéomique. Il n'y a pas de définition « officielle » du mot protéomique. Dans un sens large

elle correspond à l'ensemble des techniques permettant d'analyser les protéines. Le mot protéome a été créé en 1994 pour désigner l'ensemble des protéines d'une cellule à un instant donné (comme le génome désigne l'ensemble de son patrimoine génétique, ou le transcriptome l'ensemble de ses transcrits) et résulte de la contraction de : protéines exprimées par un génome.

Si l'on dispose d'un anticorps dirigé contre l'un des partenaires, il est possible d'immunoprécipiter les complexes de protéines en interaction avec cet anticorps : c'est la co-immunoprécipitation. Le problème majeur est qu'en général les complexes co-précipités contiennent de nombreux contaminants dont il est difficile de se débarrasser. Une approche qui permet d'obtenir des complexes mieux purifiés fait appel aux étiquettes : histidine (tag-His), glutathion S-transférase (tag-GST) ou Flag (tag-Flag) par exemple. Pour cela l'un des partenaires, qui sert d'appât, est produit par une bactérie avec une étiquette (His, GST, Flag,…) fixée à son extrémité N-terminale ou C-terminale. Cette étiquette permet dans un premier temps sa purification. Il est ensuite mis en contact avec un extrait cellulaire afin qu'il interagisse avec ses partenaires. Les complexes formés durant l'incubation sont purifiés par affinité (sur une colonne de Ni^{2+} pour un tag-His, de glutathion pour un tag-GST, colonne d'anticorps anti-Flag pour le tag-Flag). Quelle que soit la méthode utilisée pour purifier les complexes, leur composition peut être analysée grâce aux outils de la protéomique.

En utilisant les GFP il est possible de caractériser des interactions protéine/protéine *in vivo*. Il faut pour cela coupler chacune des deux protéines dont on souhaite étudier l'interaction à des GFP de couleurs différentes. La longueur d'onde d'excitation d'une des deux GFP doit correspondre à la longueur d'onde d'émission de l'autre, par exemple des GFP de couleur cyan et jaune. Si les deux GFP sont suffisamment proches l'une de l'autre (ce qui indique que les protéines auxquelles elles sont couplées sont suffisamment proches, c'est- à-dire interagissent), il se produit un phénomène de FRET (*Fluorescence Resonance Energy Transfert*). La lumière émise par la première GFP (cyan) lorsqu'elle est excitée (il faut un rayonnement monochromatique qui n'excite pas en même temps la GFP jaune) est captée par l'autre GFP qui émet alors une lumière jaune. Si les deux protéines couplées aux deux GFP sont trop éloignées, le phénomène de FRET ne peut se produire et il n'y a pas émission de lumière jaune.

D'autres techniques plus lourdes existent, notamment la spectrométrie de masse (technique physico-chimique qui permet d'identifier la séquence en acides aminés d'une protéine). Cependant, ce ne sont en général pas des techniques réalisées en premier abord dans le cas des facteurs de transcription, il faut déjà avoir étudié la protéine et avoir une idée sur ses partenaires.

D. Mise en évidence de l'action biologique des séquences régulatrices

D'après (Kaplan et Delpech, 2007).

Les expériences décrites dans le paragraphe précédent ne sont en fait qu'une étape dans la recherche de séquences régulatrices (qui interagissent le plus souvent avec des facteurs de transcription), car rien ne nous dit que les séquences trouvées possèdent un rôle biologique réel : il peut y avoir des interactions non attendues ou des biais (une séquence peut être sélectionnée à tort). Il est donc nécessaire de valider l'action biologique des séquences potentiellement régulatrices trouvées. Pour cela, plusieurs stratégies existent et ont été utilisées par différents auteurs mais il n'existe pas de technique générale permettant d'aboutir à la démonstration de ce rôle biologique : il faut en fait combiner certaines des techniques déjà décrites avec celles qui vont être décrites dans ce chapitre. L'un des problèmes majeurs est que, pour être significatives, les études doivent être conduites *ex vivo* (cellules en culture) ou *in vivo* (modèle souris par exemple).

1. Constructions permettant de mettre en évidence les effets des modifications apportées

Chaque fois que l'on veut analyser les effets en cis d'une séquence, ou les effets sur une séquence d'un facteur de transcription, le plus simple est d'avoir recours à la technique du gène rapporteur. L'expérimentation devant être conduite dans des cellules vivantes, la première étape consiste en la construction d'un vecteur selon le schéma de base suivant :

- la séquence potentiellement régulatrice, dont on souhaite étudier l'effet sur le taux de transcription, doit être couplée à un gène, non exprimé dans la cellule hôte, et dont le produit est facilement analysable et quantifiable, que l'on appelle rapporteur (ou *reporter*). Du fait de la construction, son expression se trouve alors sous le contrôle de la séquence qui lui a été greffée.

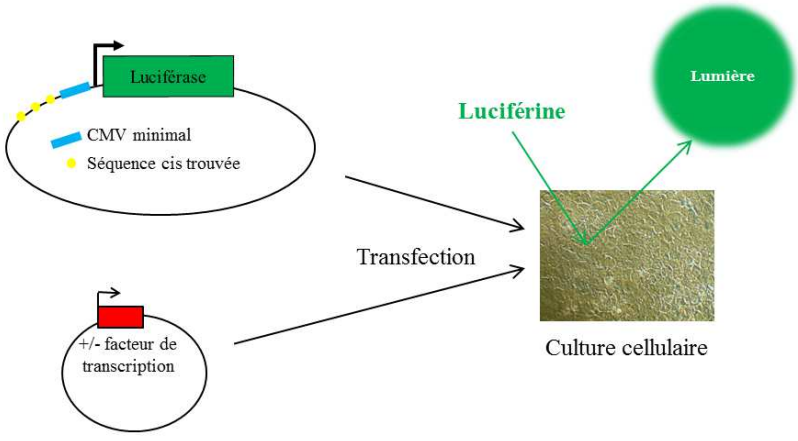
- un marqueur doit être ajouté pour permettre ultérieurement de sélectionner les cellules qui ont intégré le vecteur.

- il s'agit là d'un vecteur minimum, le plus souvent la réponse aux questions posées nécessite l'apport d'autres séquences spécifiques.

- un plasmide permettant l'expression du facteur de transcription d'intérêt.

La construction est ensuite introduite dans des cellules en culture (transfection ou infection virale) ou plus rarement dans un ovocyte fécondé ou des cellules ES (animaux transgéniques). Les cellules qui ont incorporé le vecteur sont sélectionnées. L'expression du gène rapporteur est ensuite analysée : le taux de transcription du gène rapporteur ou même le produit protéique de ce gène serviront à mesurer l'effet à la fois de la séquence cis incorporée et à la fois du ou des facteurs de transcription ajoutés (Figure 30).

Figure 30 : Essai luciférase



Une expérience classique d'essai luciférase : on réalise des vecteurs recombinants comportant un gène rapporteur couplé aux séquences cis, cibles de facteurs de transcription dont on souhaite étudier l'effet. La construction est transfectée dans des cellules et le taux d'expression du gène rapporteur est mesuré.

En pratique cette expérience réalisée de manière isolée n'a que peu d'intérêt. Il convient de construire toute une série de vecteurs dans lesquels la séquence potentiellement régulatrice (promoteur, *enhancer, silencer*) aura été mutée, partiellement délétée ou déplacée par rapport au gène rapporteur. La comparaison des taux d'expression du gène rapporteur dans ces différents vecteurs permet de définir les séquences et les bases directement impliquées dans la régulation, mettant ainsi en évidence un éventuel effet de position.

Les gènes rapporteurs les plus utilisés sont : le gène de la chloramphénicol acétyl-transférase (CAT), le gène de la bêta-galactosidase, le gène de la bêta-glycuronidase, et le gène de la luciférase (il est aussi possible de faire appel aux GFP – *Green Fluorescent Protein* – mais elles sont surtout utilisées pour repérer la localisation d'une protéine au sein d'une cellule, plutôt que pour quantifier l'expression d'un gène). Le gène de la bêta-galactosidase ou de la luciférase ou encore d'une GFP présente l'avantage de permettre une analyse *in situ*. En effet la bêta-galactosidase peut métaboliser un composé artificiel, le X-gal (ou ses équivalents), qui acquiert une coloration bleue lorsqu'il est métabolisé. La luciférase est une enzyme extraite du ver luisant qui induit l'émission de lumière en présence d'ATP et de son substrat, la luciférine. Ces activités peuvent donc être parfaitement observées avec un microscope. Cette possibilité présente un intérêt particulier lorsque seules certaines cellules répondent. Des résultats particulièrement spectaculaires ont été obtenus au cours des études sur le développement où ce type d'expériences a permis d'étudier la localisation et la cinétique de l'expression de gènes du développement.

Toutes ces expériences, simples dans leur principe, sont d'exécution délicate et nécessitent pour être significatives un grand nombre d'expériences de contrôle.

2. Études transcriptomiques

Par définition nous l'avons vu, un facteur de transcription est capable de moduler la transcription de gènes-cibles. Indirectement donc, nous pouvons étudier les effets sur la transcription d'une surexpression ou d'un KO (ou de toute autre mutation) d'un facteur de transcription dans un modèle cellulaire en étudiant le transcriptome de ces cellules, c'est-à-dire les quantités et les proportions des différents transcrits, par PCR quantitative ou à l'aide de puces à ARN (techniques que je ne détaillerai pas dans cette thèse).

3. Les modèles biologiques animaux (transgénèse, KO, KI, KO conditionnel)

Une approche très puissante lorsqu'un nouveau facteur de transcription a été caractérisé consiste à le faire exprimer par une cellule ou un animal (qui la possède déjà ou qui ne la possède pas). Cela nécessite un travail considérable et dont il ne peut être déterminé à l'avance s'il sera couronné de succès. Lorsqu'un nouveau facteur de transcription a été caractérisé et a montré son importance (en utilisant les techniques évoquées plus haut), le modèle animal est le meilleur moyen de connaître toutes les actions biologiques de ce facteur, dans tous les organes, et d'envisager possiblement des essais thérapeutiques (si le facteur est impliqué dans une maladie particulière).

Les différentes techniques d'intégration sont, brièvement (Kaplan et Delpech, 2007) :

- la transgénèse additive (avec ou sans promoteur inductible) : le gène que l'on intègre peut être un gène artificiellement muté (étude de physiopathologie moléculaire) ou un gène non exprimé dans la cellule ou l'animal (démonstration de la capacité d'une protéine à induire une différenciation), ou un gène normal que l'on souhaite faire surexprimer. Les promoteurs inductibles sont des promoteurs activables à la demande par une molécule donnée. Ce type de promoteur va permettre de déclencher à la demande l'expression d'un gène donné afin de déterminer ses effets sur la cellule ou l'organisme. Il y a quelques années les plus utilisés étaient une séquence du LTR

(*Long Terminal Repeat*) du virus de tumeur mammaire de la souris (MMTV, *Mouse Mammary Tumor Virus*) qui est activable par la dexaméthasone, et le promoteur du gène de la métallothionéine inductible par la dexaméthasone et les ions zinc ou cadmium. Les problèmes de ces promoteurs étaient que l'induction nécessitait plusieurs heures, avec un taux d'induction faible, et parfois une activation du gène sans même apporter l'inducteur. Cela a conduit à développer des systèmes beaucoup plus performants, le système d'induction ou de répression par les tétracyclines (dit Tet-on et Tet-off) et le système d'induction par un analogue des œstrogènes.

- la souris invalidée (*knock-out*, KO) ou plutôt la souris KO conditionnel : en effet, nous avons vu que les mutations dans les facteurs de transcription étaient souvent délétères ; l'inactivation d'un facteur de transcription chez la souris a donc de forte chance d'entraîner la mort *in utero* (souris non viables). Il est donc intéressant dans ce cas d'inactiver le gène à un moment précis, ou dans un tissu précis, une fois le développement embryonnaire terminé. Il existe pour cela de systèmes cre-lox inductible.

- la souris KI, *knock-in* (plus rare), qui consiste à introduire une mutation de façon ciblée.

DEUXIÈME PARTIE : TRAVAIL PERSONNEL

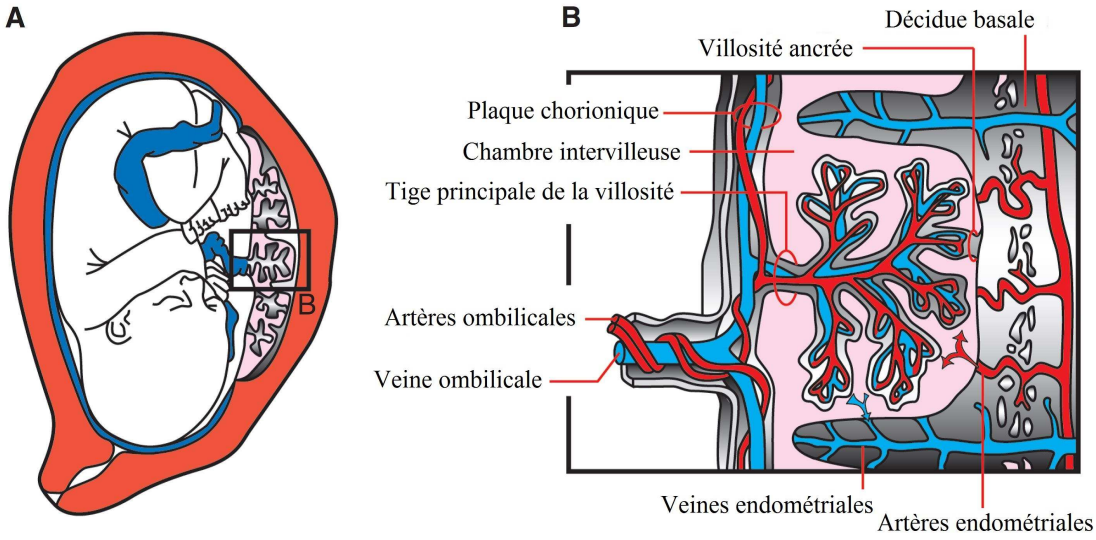
IDENTIFICATION DU SITE CONSENSUS DE FIXATION DU FACTEUR DE TRANSCRIPTION STOX1

I. Introduction – contexte

Dans les pays industrialisés, une des principales maladies de la grossesse est la prééclampsie. Il s'agit d'une maladie syndromique et multifactorielle, spécifiquement humaine, qui affecte entre 3 et 8 % des grossesses. Elle est caractérisée par une hypertension gestationnelle *de novo* (Pression Artérielle Systolique ≥ 140 mmHg et Pression Artérielle Diastolique ≥ 90 mmHg) et une protéinurie (> 300 mg/24 heures), se développant à partir de la 20ème semaine d'aménorrhée. Soixante-quinze pour cent des prééclampsies se développent lors de la première grossesse, ce qui rend cette maladie difficile à prévoir. Les symptômes de la maladie s'aggravent tout au long de la grossesse, et peuvent potentiellement entraîner le décès de la mère (environ 20 décès par an en France). À ce jour, les solutions pharmaceutiques préventives ou thérapeutiques sont très limitées, et le seul moyen efficace de stopper la progression de la prééclampsie reste le retrait de l'unité fœto-placentaire. Ainsi, la prééclampsie est responsable d'environ 15 % des naissances prématurées dans les pays occidentaux. Le défi majeur de l'obstétricien est donc de permettre à la grossesse de se poursuivre, afin d'améliorer le pronostic pour l'enfant, sans porter atteinte à la santé de la mère (d'après le site web du CNGOF, Collège National des Gynécologues et Obstétriciens Français : http://www.cngof.asso.fr/index.html).

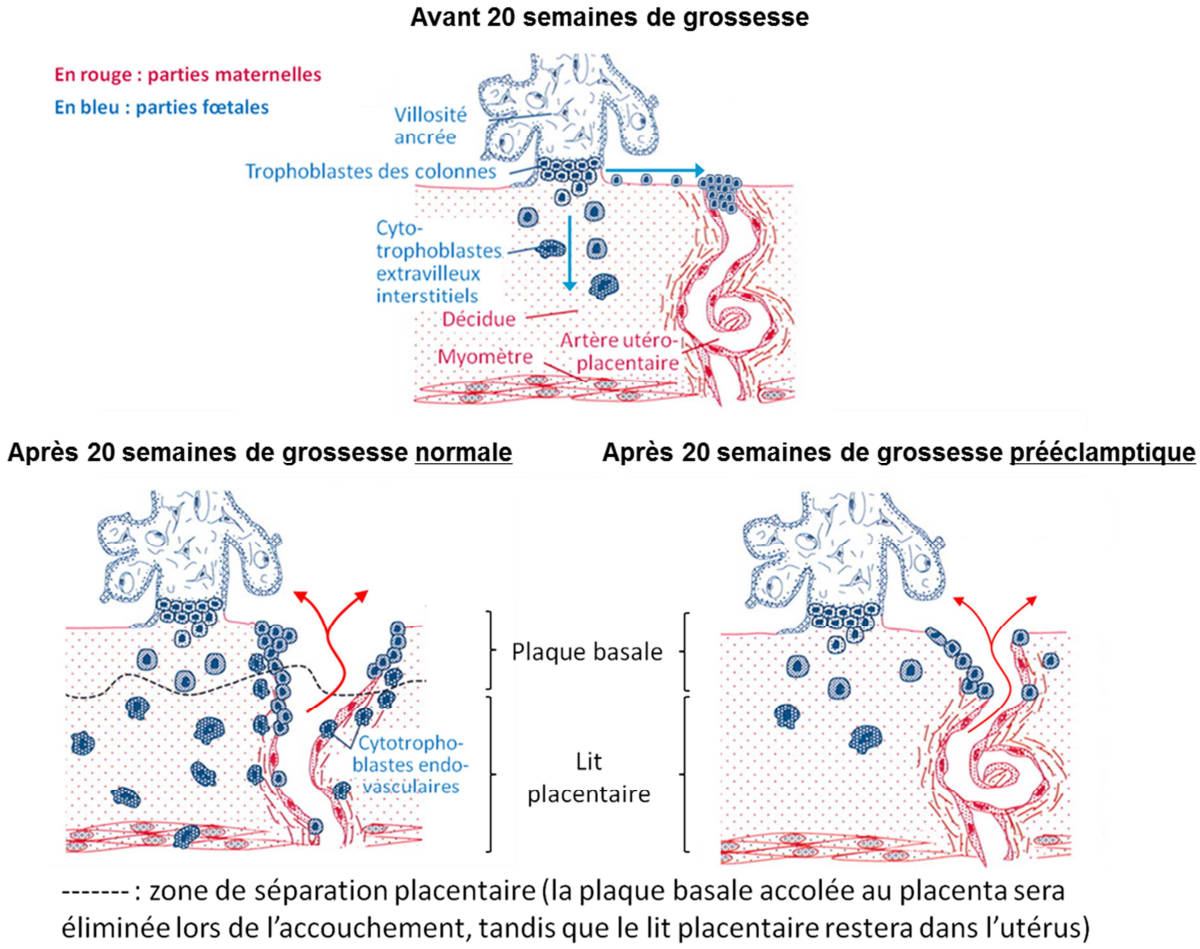
Même s'il semble établi que le placenta soit à l'origine de cette maladie, les mécanismes impliqués dans le développement du syndrome maternel restent mal connus. Le consensus admis aujourd'hui est qu'un défaut de placentation (Figure 31) entraînerait une mauvaise vascularisation placentaire (et donc une moins bonne perfusion) et qu'en parallèle, le placenta libèrerait des substances vaso-actives qui altéreraient l'endothélium vasculaire maternel, entrainant l'hypertension maternelle (Figure 32).

Figure 31 : Schéma d'un placenta humain



D'après (Fuchs et Ellinger, 2004).
A : Position du fœtus et du placenta dans l'utérus.
B : Agrandissement d'une unité fœto-placentaire. Le placenta humain est un placenta de type histologique hémochorial : c'est le type de placenta chez les mammifères où les sangs maternel et fœtal sont séparés par le moins de couches cellulaires. Des artères endométriales (ou artères spiralées utérines) déversent du sang riche en oxygène dans une chambre intervilleuse. Les échanges (oxygène, CO_2, acides aminés,…) se font au niveau de villosités qui contiennent de multiples vaisseaux sanguins fœtaux. Ces vaisseaux sont séparés du sang de la mère par leur endothélium et une couche de syncytiotrophoblastes (couche de cellules fœtales ayant fusionnées). Le sang pauvre en oxygène dans la chambre intervilleuse regagne la circulation maternelle via des veines endométriales.

Figure 32 : Le défaut de remodelage des artères spiralées utérines en cas de prééclampsie



Adapté de (Kaufmann *et al.*, 2003). Au cours de la formation du placenta humain, une sous-population de trophoblastes, les cytotrophoblastes extravilleux, envahissent l'endomètre utérin jusqu'à atteindre les artères utéro-placentaires où ils prennent la place des cellules endothéliales maternelles. Lors d'une grossesse normale, ce remodelage entraîne la perte de la capacité de réponse des artères spiralées à des substances vaso-actives, ce qui se traduit par un élargissement fonctionnel des artères et une perfusion sanguine adéquate au bon développement du fœtus. En cas de prééclampsie, ce remodelage se ferait mal, entraînant une moins bonne perfusion. L'hypertension artérielle maternelle observée dans ce syndrome permet probablement de compenser cette moins bonne perfusion. Cependant, les mécanismes menant à cette hypertension ne sont pas encore parfaitement élucidés. Des éléments de réponses ont tout de même émergé au cours de ces dix dernières années, notamment grâce à la description de la libération par le placenta prééclamptique de substances spécifiquement impliqués dans l'angiogenèse (comme les récepteurs solubles du VEGF – *Vascular endothelial growth factor* – et de l'Endogline, respectivement sFlt1 et sEng) qui participent à l'établissement de l'hypertension.

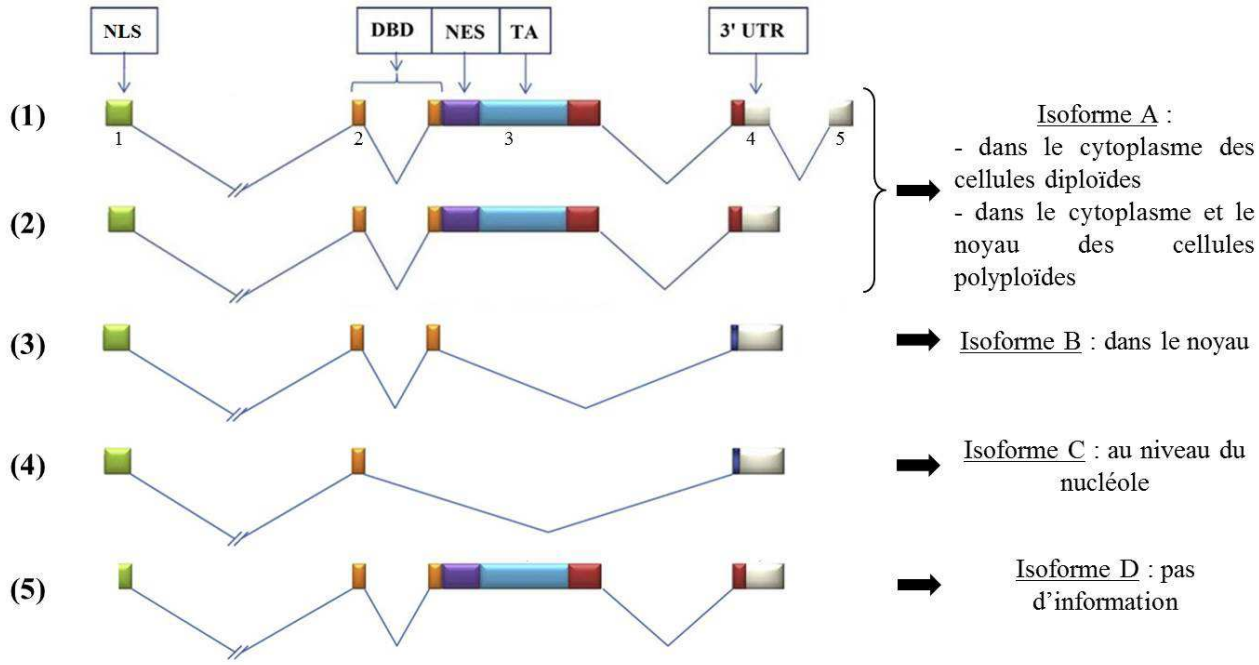
La prééclampsie reste difficile à étudier directement dans l'espèce humaine car nous n'avons accès qu'à des placentas à un terme avancé de la maladie quand les causes et conséquences de la maladie sur l'unité fœto-maternelle sont indistinguables. De plus, jusqu'à peu, le manque de modèles animaux limitait l'investigation de l'étiopathogénie de cette maladie et l'évaluation de nouveaux médicaments. Cependant, un nouveau modèle murin de prééclampsie sévère a récemment été développé au sein du laboratoire du docteur Vaiman (Doridot *et al.*, 2013). Ce modèle, que je détaillerai par la suite, représente une piste prometteuse dans l'exploration de voies physiopathologiques de la prééclampsie.

Malgré les lacunes dans la compréhension de la physiopathologie de la prééclampsie, cette maladie est connue comme étant un syndrome multifactoriel qui a une base génétique polygénique, et une héritabilité estimée à environ 55 % à partir de l'analyse de grandes cohortes scandinaves (Cnattingius *et al.*, 2004). *STOX1* (*Storkhead box 1*) est le premier gène dont un variant a été associé à cette maladie (van Dijk *et al.*, 2005). Plusieurs études ont par la suite montré son implication dans la prééclampsie. En 2009, une analyse du transcriptome de biopsies placentaires du premier trimestre de la grossesse a révélé que l'expression de *STOX1* était significativement augmentée (2,1 fois) dans le placenta des femmes qui vont développer une prééclampsie (Founds *et al.*, 2009). De plus, dans des modèles cellulaires, il a été montré que *STOX1* modulait l'invasion cytotrophoblastique (van Dijk *et al.*, 2005) et que la surexpression du gène dans une lignée de choriocarcinome humain (JEG-3) entraînait des altérations transcriptomiques très significativement corrélées à celles observées dans des placentas prééclamptiques (Rigourd *et al.*, 2008). Ainsi, la surexpression précoce de *STOX1* serait une caractéristique d'une placentation anormale et serait liée à la prééclampsie.

L'équipe du Docteur Daniel Vaiman à l'Institut Cochin, qui s'intitule « Génomique, Épigénétique et Physiopathologie de la Reproduction », s'intéresse entre autres au rôle de *STOX1* dans la physiopathologie de la prééclampsie. Initialement basé sur des modèles cellulaires et des matériels biologiques humains (ARNs placentaires, ADNs de patientes, plasmas de patientes), ce travail a pris un nouveau tournant depuis que le laboratoire a développé un modèle de souris transgénique (avec un fond génétique FVB/N) par transgénèse additive de *STOX1*, afin d'étudier son impact sur la placentation. Lorsque des souris femelles sauvages ont été croisées avec des mâles transgéniques (ce qui permet la restriction de l'expression du transgène à la sphère fœto-placentaire), ces souris ont développé un syndrome prééclamptique sévère, en dépit du fait que la prééclampsie ne se développe pas spontanément chez les rongeurs. En effet, au cours et seulement pendant la gestation, ces souris présentaient une hypertension sévère (Pression Artérielle Systolique > 160 mmHg), une protéinurie ainsi qu'une élévation de marqueurs sériques de la prééclampsie (sFlt1 et sEng). Les modifications physiopathologiques observées chez les souris surexprimant *STOX1* dans l'unité fœto-placentaire étant semblables à celles de la prééclampsie humaine, il est fort probable que des mécanismes à l'origine de l'établissement de ce syndrome soient communs. Ce modèle est à notre connaissance et à ce jour le seul modèle de prééclampsie sévère chez les rongeurs, et constitue un atout unique pour déchiffrer de nouvelles voies impliquées dans la physiopathologie de la prééclampsie, et pour tester de nouvelles approches thérapeutiques.

Si le rôle du gène *STOX1* dans la prééclampsie est donc maintenant bien établi, on connaît cependant peu de choses sur la biochimie des protéines qu'il code. En effet, d'après les bases de données NCBI (*National Center for Biotechnology Information* : http://www.ncbi.nlm.nih.gov.gate2.inist.fr/) et *Ensembl* (http://www.ensembl.org/), cinq transcrits différents sont générés et codent quatre isoformes (A, B, C et D, voir Figure 33 et Annexe 3). L'isoforme A a été défini dans le papier princeps (van Dijk *et al.*, 2005) comme un facteur de transcription. Les auteurs ont utilisé trois logiciels en ligne : le programme SIFT (*Sorting Intolerant From Tolerant* : http://blocks.fhcrc.org/sift/sift.html) pour prédire si une substitution d'acide aminé aurait une incidence sur la fonction de la protéine ; le programme PSIPRED (*Protein Structure Prediction server* : http://bioinf.cs.ucl.ac.uk/psipred/) pour l'analyse de structure secondaire ; et le programme MotifScan (http://scansite.mit.edu) pour identifier les motifs protéiques. De plus, la comparaison de sa séquence avec d'autres protéines par analyse bio-informatique (Annexe 4) a permis de prédire que l'isoforme A, qui inclut l'ensemble des exons, contient un signal de localisation nucléaire, un domaine de fixation à l'ADN contenant des acides aminés chargés majoritairement positivement et se rapprochant de celui de la famille des facteurs transcriptionnels FOX (*Forkhead box*, Annexe 4), et un domaine de transactivation (Figure 33). Cependant, on ne connait ni les séquences avec lesquelles il interagit, ni quels potentiels gènes il régule directement.

Figure 33 : Transcrits et isoformes de STOX1



D'après (Rigourd *et al.*, 2009). Le gène *STOX1* code quatre isoformes: A (989 acides aminés), B (227 acides aminés), C (169 acides aminés) et D (879 acides aminés). Ces isoformes sont situées dans différents compartiments cellulaires dans les cellules placentaires. La localisation cellulaire de l'isoforme A (noyau ou cytoplasme) dépend de l'état de ploïdie cellulaire (van Dijk *et al.*, 2005). Les isoformes B et C sont situées dans le noyau. L'ARNm qui code l'isoforme D est très peu transcrit et il n'y a eu à ce jour aucune étude sur sa localisation cellulaire. Les couleurs identiques à l'intérieur des exons indiquent les séquences de protéines identiques.
NLS (*Nuclear Localization Sequence*) : Séquence de localisation nucléaire.
DBD (*DNA Binding Domain*) : Domaine de liaison à l'ADN.
NES (*Nuclear Export Signal*) : Signal d'export nucléaire.
TA (*Transactivator domain*) : Domaine transactivateur.
Par épissage intra-exonique, le domaine transactivateur dans l'isoforme B est supprimé. Par épissage alternatif de l'exon 3, une partie du DBD dans l'isoforme C est enlevée. Les différentes couleurs des derniers exons indiquent que le site d'épissage est différent d'un isoforme à l'autre (voir Annexe 3).

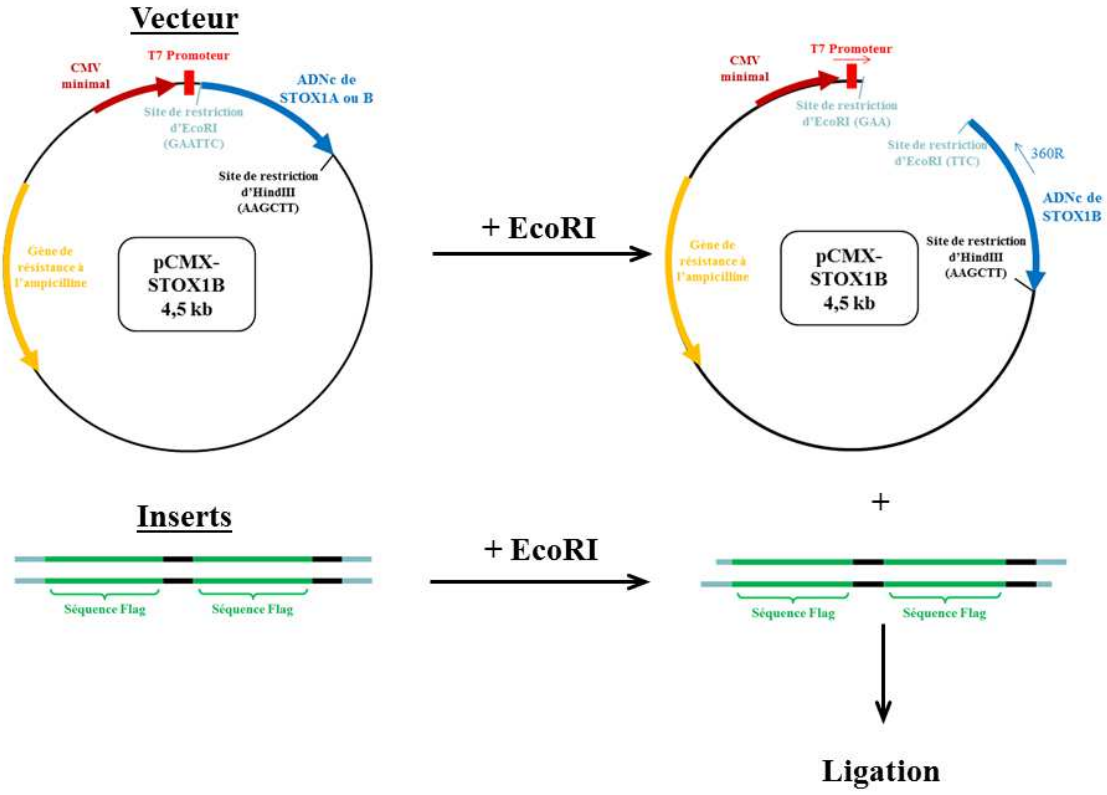
J'ai effectué un stage de recherche dans le laboratoire du docteur Daniel Vaiman, dont le but était d'identifier le site de fixation à l'ADN de la protéine STOX1 via la méthode de PCR-sélection, dont le principe est de purifier des oligonucléotides fixés par un facteur transcriptionnel sélectionnés par celui-ci au sein d'une banque d'oligonucléotides aléatoires. Pour cela, j'ai d'abord construit un vecteur plasmidique permettant l'expression d'une protéine chimérique Flag-STOX1 pour les isoformes A et B (qui contiennent toutes les deux le domaine présumé de liaison à l'ADN), dans le but final d'utiliser un anticorps anti-Flag très spécifique lors de la phase de purification. Cette technique m'a permis de déterminer deux séquences consensus possibles du site de fixation qui sont en cours de validation via des expériences de retard sur gel et d'essais luciférase.

II. Matériel et méthodes

A. Construction d'un vecteur d'expression pour Flag-STOX1A et Flag-STOX1B

Pour produire notre protéine chimérique Flag-STOX1 (isoforme A ou B), nous avons utilisé un vecteur déjà disponible au laboratoire, où la séquence d'ADNc de *STOX1* (isoforme A ou B) humain avait été intégrée dans le vecteur pCMX entre les sites des enzymes de restriction EcoRI et HindIII (Figure 34). Le Flag en lui-même correspond à la séquence protéique DYKDDDDK (Acide aspartique – Tyrosine – Lysine – Acide aspartique – Acide aspartique – Acide aspartique – Acide aspartique – Lysine). Pour construire l'insert, j'ai commandé des oligonucléotides contenant deux séquences de Flag encadrées par les séquences reconnues par l'enzyme de restriction EcoRI (GAATTC) : GAA TTC ATG GAC TAC AAG GAC GAC GAT GAC AAG GGA TCC GAC TAC AAG GAC GAC GAT GAC AAG GGA TCC GAA TTC (75 pb). Le vecteur ainsi que des inserts Flag ont été digéré avec l'enzyme de restriction EcoRI de chez Life Technologies™ (Carlsbad, États-Unis). Le vecteur a ensuite été déphosphorylé avec de la SAP (*Shrimp Alkaline Phosphatase*) de chez Affymetrix® (Santa Clara, États-Unis) : cette étape est essentielle pour éviter la recircularisation du plasmide. Le vecteur déphosphorylé a ensuite été purifié sur colonne avec le kit « NucleoSpin® Gel and PCR Clean-up » de chez Macherey-Nagel (Düren, Allemagne) avant ligation. La ligation a été réalisée avec de la ligase T4 de chez Thermo Scientific (Waltham, États-Unis) toute une nuit en présence du vecteur digéré purifié et des inserts digérés purifiés présents en excès molaire (5 à 10 fois). Des bactéries chimiocompétentes (« One Shot® TOP10 Chemically Competent E. coli », Life Technologies™) ont ensuite été transformées avec le mélange de ligation et étalées sur milieu sélectif (LB-Ampicilline). Après une nuit à 37°C, une PCR a été réalisée sur les clones obtenus avec des amorces situées de part et d'autre des inserts attendus (T7P et 360R, Figure 34). Cela permet de voir, grâce à une migration sur gel d'agarose à 2 %, lesquelles ont intégré un insert ou éventuellement des concatémères d'inserts. Les colonies ayant intégré un ou plusieurs inserts ont alors été mises en culture (milieu LB + ampicilline 100 µg/mL) toute une nuit. Le lendemain, l'ADN plasmidique de ces colonies a été extrait par la technique « miniprep » grâce au kit « GeneJET™ Plasmid miniprep Kit » de chez Thermo Scientific et séquencé à la plate-forme de séquençage de l'Institut Cochin afin de vérifier le nombre d'inserts correctement intégrés et leur sens.

Figure 34 : Construction du vecteur Flag-STOX1



Pour produire notre protéine chimérique Flag-STOX1 (isoforme A ou B), nous avons utilisé un vecteur déjà disponible au laboratoire, où la séquence d'ADNc de *STOX1* (isoforme A ou B) humain a été intégrée dans le vecteur pCMX entre les sites des enzymes de restriction EcoRI et HindIII. Le Flag en lui-même correspond à la séquence protéique DYKDDDDK (Acide aspartique – Tyrosine – Lysine – Acide aspartique – Acide aspartique – Acide aspartique – Acide aspartique – Lysine). Pour construire l'insert, j'ai commandé des oligonucléotides contenant deux séquences de Flag encadrées par les séquences reconnues par l'enzyme de restriction EcoRI (GAATTC) : GAA TTC ATG GAC TAC AAG GAC GAC GAT GAC AAG GGA TCC GAC TAC AAG GAC GAC GAT GAC AAG GGA TCC GAA TTC (75 pb). Le vecteur ainsi que des inserts Flag ont été digéré avec l'enzyme de restriction EcoRI. Le vecteur a ensuite été déphosphorylé avec de la SAP (*Shrimp Alkaline Phosphatase*) pour éviter la recircularisation du plasmide. Le vecteur déphosphorylé a ensuite été purifié sur colonne avant ligation.
CMV minimal (Cytomégalovirus) : séquence promotrice forte permettant l'initiation de la transcription (contenant notamment la boîte TATA, la boîte CAAT,…)

B. Transfection dans des cellules COS et extraction protéique

Les cellules COS-7 ont été cultivées dans un milieu complet DMEM(1X)+GlutaMAX™-1 (Life Technologies™) contenant 10 % de sérum de veau fœtal (SVF) et 1 % de pénicilline-streptomycine. L'ensemencement a été effectué entre 35 et 45 % de confluence la veille de la transfection (de façon à être en phase exponentielle le lendemain au moment de la transfection). Les cellules ont été transfectées avec la lipofectamine®2 000 de chez Life Technologies™ qui contient des liposomes pouvant piéger les molécules d'ADN (les liposomes, en traversant la bicouche lipidique des membranes des cellules eucaryotes, peuvent ainsi faire rentrer des molécules d'ADN dans les cellules). Pour chaque transfection dans un puits d'une plaque 6 puits de chez TPP® (Trasadingen, Suisse), 2 µg d'ADN ont été mélangé dans 200 µL de DMEM simple (sans SVF ni antibiotique) contenant 6 µL de lipofectamine. Après 6 heures d'incubation, du SVF a été rajouté afin d'en avoir 10 % dans le milieu. Pour certaines transfections, du MG132 réf. C2211 de chez Sigma-Aldrich® (Saint-Louis, États-Unis) a été rajouté : il s'agit d'un inhibiteur du protéasome

devant être utilisé à 20 µM dans le milieu de culture pendant 4 heures avant l'extraction protéique. L'extraction protéique en elle-même a eu lieu 48 heures après la transfection : les cellules ont été décollées des puits par grattage puis récupérées dans un tampon d'extraction RIPA contenant des anti-protéases (25 mM NaCl + 10 mM Tris-HCl pH7,5 + 5 mM EDTA + 0,1% NP-40 + 1X Protease Inhibitor Cocktail de chez Sigma). Les cellules ont ensuite été soniquées 5 fois 5 secondes dans le sonificateur Bioruptor® Standard de chez Diagenode (Liège, Belgique), puis laissées 1 heure sous agitation à 4 °C. Après centrifugation pour sédimenter les débris cellulaires (5 minutes à 13 000 g), le surnageant contenant les extraits protéiques a été prélevé. Une partie de ce surnageant a été conservée pour doser les protéines (par la méthode de Bradford), une autre partie pour le Western blot (WB) de contrôle. Une fois le WB de contrôle réalisé, une transfection à plus grande échelle a été réalisée, en vue de l'expérience de PCR-sélection, dans des boîtes rondes de culture cellulaire de 60,1 cm² (TPP®). Les quantités de milieu, de plasmide et de lipofectamine dans ces boîtes ont été calculées au *prorata* de la surface cellulaire (un puits d'une plaque 6 puits faisant 8,960 cm²).

C. Western blot : vérification de l'expression de la protéine

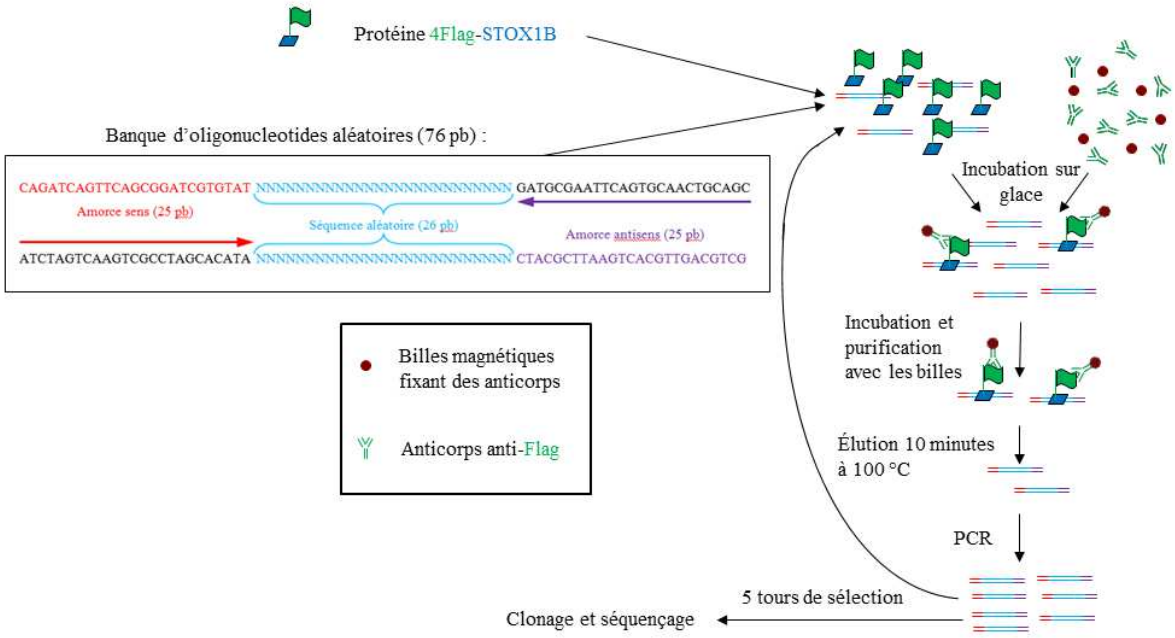
Une partie de l'extrait protéique obtenu a été mélangée à un tampon de dénaturation Laemmli contenant du bêta-mercapto-éthanol (50 mM Tris pH6,8 + 2 % SDS + 10 % glycérol + 0,004 % bleu de bromophénol + 5 % bêta-mercapto-éthanol) de façon à déposer 10 à 30 µg de protéine. Les échantillons ont ensuite été chauffés à 100 °C pendant 5 minutes afin de les dénaturer, puis déposés sur un gel de polyacrylamide en condition dénaturante (SDS-PAGE). Le gel de concentration (qui permet de concentrer les échantillons protéiques sur la même ligne de départ avant séparation) et le gel de séparation 10 % (qui permet de séparer les protéines en fonction de leur poids moléculaire apparent) ont été coulés (voir Annexe 6 pour la composition). Après migration des protéines dans un tampon adéquat (Tris 25 mM + Glycine 190 mM + SDS 2 %) puis transfert sur une membrane de PolyVinyliDene Fluorède (PVDF) Amersham Hybon^{TM}-P de chez GE Healthcare (Pittsburgh, États-Unis) dans un autre tampon adéquat (glycine 192 mM + Tris base 25 mM pH8,3), les sites aspécifiques de la membrane ont été bloqués par les protéines de lait dans un tampon de blocage de PBS contenant 5 % de lait en poudre. Puis les membranes ont été incubées une nuit à 4 °C avec l'anticorps primaire anti-Flag M2 (Sigma F3165) dilué au 2 500ème dans un tampon de PBS contenant 0,1 % de tween et 1 % de lait. Le lendemain, après trois lavages des membranes dans du PBS contenant 0,1 % de tween, l'anticorps secondaire anti-souris (DAKO P0260 de chez Sigma) dilué au 2 500ème dans de la gélatine a été ajouté pendant environ une heure à température ambiante. Cet anticorps est couplé à la peroxydase de raifort (*Horseradish peroxydase*, HRP). Cette peroxydase catalyse la production d'H_2O et O^- à partir de H_2O_2 et l'oxydation successive du luminol qui se retrouve alors dans un état excité et émet de la lumière autour de 430 nm (indétectable à l'œil nu). La détection a été réalisée grâce au kit de chimioluminescence « ECL SuperSignal® West Pico Chemiluminescent Substrate » de chez Thermo Scientific, et la révélation sur un film photographique.

D. PCR-sélection

Le mélange suivant a été incubés pendant 40 minutes : 50 µg de protéines issues des extraits protéiques des transfections (qui étaient donc conservées dans du tampon RIPA : 25 mM NaCl + 10 mM Tris-HCl pH7,5 + 5 mM EDTA + 0,1% NP-40 + 1X Protease Inhibitor Cocktail de chez Sigma), en présence de 5,5 µg de poly(dI-dC) (ou acide Poly(deoxyinosinic-deoxycytidylic), un copolymère double brin aléatoire ne contenant que des bases Cytidine ou Inosine, qui, en entrant en compétition avec l'ADN, diminue les fixation non-spécifiques entre l'ADN et les protéines), avec 3 ng d'une solution d'oligonucléotides aléatoires double brin de 76 pb (construits avec deux séquences fixes de 25 pb – qui ont été utilisées pour les amorces de PCR – encadrant une séquence aléatoire de 26 pb : Figure 31), dans un volume total de 50 µL d'eau exempte de DNAse et RNAse.

En parallèle, 1,5 mg de billes Dynabeads® Protein G de chez Life Technologies™ ont été lavées deux fois avec 400 µL de PBS. Ces billes ont été incubées pendant 20 minutes sous agitation avec 1 µg de l'anticorps anti-Flag M2 dans 200 µL de PBS-0,5 % NP40. Les billes couplées aux anticorps ont ensuite été incubées sous agitation pendant 20 minutes avec le mélange contenant les extraits protéiques et les oligonucléotides aléatoires. Après trois lavages avec 400 µL de PBS, les complexes billes-anticorps-protéines-ADN ont été resuspendus dans 30 µL d'eau puis dénaturés 10 minutes à 100 °C. Le surnageant a ensuite été récupéré et on a réalisé des PCR (grâce aux amorces des oligonucléotides) avec différents nombres de cycle (Figure 36) dont 4 µL du produit a été déposé sur un gel d'agarose 3 %, ce qui a permis de vérifier la taille et choisir le produit de PCR le plus riche en ADN spécifique en fonction du nombre de cycle. Le produit de PCR le plus adéquat a été choisi pour réaliser un autre tour de PCR-sélection, qui consiste à réaliser à nouveau la suite d'étapes purification-immunoprécipitation-élution-PCR décrit plus haut, en remplaçant les oligonucléotides aléatoires par 3 µL de produit de PCR du tour précédent. En tout, 5 étapes consécutives ont été réalisées. À la fin du cinquième tour, les produits de PCR ont été purifiés sur gel à l'aide du kit « GeneJet Gel Extraction kit » de chez Thermo Scientific. Ces extraits purifiés ont ensuite été clonés.

Figure 35 : Schéma général de la technique de PCR-sélection réalisée



Le principe de cette technique est de purifier des oligonucléotides sur lesquels un facteur transcriptionnel se fixe, au sein d'une banque d'oligonucléotides aléatoires. Des protéines et des oligonucléotides aléatoires sont d'abord mis à incuber, puis le mélange est immunoprécipité sur des billes couplées à des anticorps reconnaissant la protéine. Une élution finale nous permet d'obtenir les séquences d'ADN retenues, et une PCR nous permet de les amplifier. Cinq étapes consécutives sont réalisées. Les oligonucléotides sélectionnés par cette technique sont ensuite clonés puis séquencés.

Figure 36 : PCR réalisée à chaque tour de la PCR-sélection

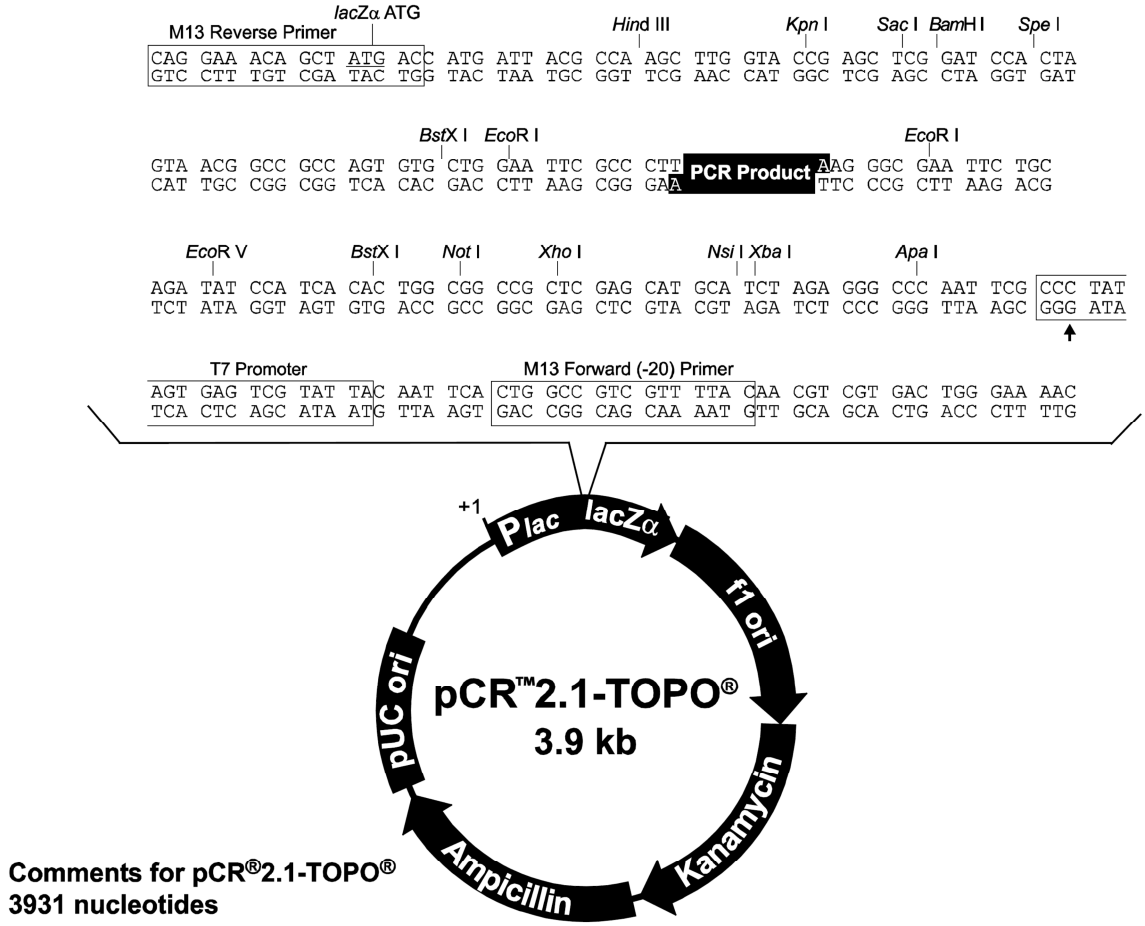
H_2O	23,8 µL	PCR : 1) 95°C : 30 sec 2) 95°C : 15 sec 3) 55°C : 15 sec 4) 72°C : 15 sec 5) 25 ou 30 cycles de 2) à 4) 6) 72°C : 1 min 7) 20°C : ∞
Tampon 5X-$MgCl_2$	10 µL	
$MgCl_2$ (25 mM)	4 µL (2 mM final)	
dNTP (10 mM)	1 µL (0,2 mM final pour chaque dNTP)	
Amorce F-76 (10 µM)	3 µL (0,6 µM final)	
Amorce R-76 (10 µM)	3 µL (0,6 µM final)	
GoTaq® (Promega)	0,2 µL	
ADN	5 µL	
Total	50 µL	

À la fin de chaque tour de PCR-sélection, le surnageant est récupéré et on réalise des PCR (grâce aux amorces des oligonucléotides) avec différents nombres de cycle dont 4 µL du produit est déposé sur un gel d'agarose à 3 %, ce qui permet de vérifier la taille et choisir le produit de PCR le plus riche en ADN spécifique en fonction du nombre de cycle. Le produit de PCR le plus adéquat est choisi pour réaliser un autre tour de PCR-sélection, qui consiste à réaliser à nouveau la suite d'étapes purification-immunoprécipitation-élution-PCR décrit plus haut, en remplaçant les oligonucléotides aléatoires par 3 µL de produit de PCR du tour précédent.

E. Clonage dans le vecteur TOPO-TA et séquençage

Le clonage dans le vecteur TOPO®TA cloning® de chez Life Technologies™ est une méthode rapide et facile pour cloner des produits de PCR et faire du séquençage. En effet, le vecteur pCR™2.1®-TOPO® de 3,9 kb ouvert est construit de telle façon à se lier facilement avec les produits de PCR qui ont eu souvent une base Adénine de rajoutée en 3' par la Taq polymérase au cours de l'élongation (Figure 37). Dans un premier temps, le produit de PCR a été incubé avec le vecteur TOPO-TA pendant 10 à 15 minutes à température ambiante (temps nécessaire pour que la topoisomérase I crée une liaison covalente entre le produit de PCR et le vecteur). Les produits de cette incubation ont été transformés dans des bactéries chimiocompétentes « One Shot® TOP10 Chemically Competent E. coli » de chez Life Technologies™, cultivées sur des boîtes de Petri Ampicilline de 10 cm de diamètre sur lesquelles on a ajouté 40 µL d'IPTG/X-gal de chez Sigma. Le lendemain, une PCR a été réalisée sur les clones obtenus dans le but de voir la taille des fragments intégrés dans le vecteur. Les colonies d'intérêt ont alors été mises en culture (milieu LB + ampicilline 100 µg/mL) toute une nuit. Le lendemain, l'ADN de ces colonies a été extrait par la technique « miniprep » grâce au kit « GeneJET™ Plasmid miniprep Kit » de chez Thermo Scientific, et séquencé à la plate-forme de séquençage de l'Institut Cochin (avec l'amorce M13F) en vue de futures analyses.

Figure 37 : Carte du vecteur pCR™2.1-TOPO®



D'après le manuel technique du kit de clonage « TOPO® TA Cloning® Kit » disponible sur le site de Life Technologies™ (http://www.lifetechnologies.com/). Le clonage dans le vecteur TOPO-TA est une méthode rapide et facile pour cloner des produits de PCR et faire du séquençage. En effet, le vecteur de 3,9 kb ouvert est construit de telle façon à se lier facilement avec les produits de PCR, qui ont eu souvent une base Adénine de rajoutée en 3' par la Taq au cours de l'élongation. Le site d'insertion est situé au milieu du gène *lacZ*, ce qui nous permet en culture bactérienne de visualiser facilement les clones ayant intégré un insert. On utilise pour cela un mélange d'IPTG et de X-gal. L'IPTG est un analogue de l'allolactose : il se lie au répresseur de l'opéron lactose et bloque son action. L'IPTG induit donc la transcription de l'opéron lactose, en particulier le gène de la bêta-galactosidase (ou *lacZ*). Le X-gal est quant à lui un galactoside (un hétéroside du galactose) : ce composé incolore peut être hydrolysé par la β-galactosidase, ce qui libère la partie indolique qui forme ensuite par oxydation un composé bleu, insoluble dans l'eau, qui précipite au site de la réaction (c'est donc ce qu'on appelle un substrat chromogénique de la β-galactosidase, c'est-à-dire qui produit une coloration lors de la réaction). Ainsi, en présence de X-gal, les bactéries qui deviennent bleues sont celles ayant intégré un plasmide où le gène *lacZ* est intact, c'est-à-dire un plasmide n'ayant pas inséré de produit de PCR.

F. Analyse bio-informatique : logiciel MEME

L'analyse des séquences a été effectuée à l'aide du logiciel MEME (*Multiple Em for Motif Elicitation*) version 4.9.0, qui est un logiciel accessible en ligne (http://meme.nbcr.net/), décrit pour la première fois dans un article en 1994 (Bailey et Elkan, 1994), et qui permet de trouver un motif commun au sein d'un groupe de séquences d'ADN ou de séquences de protéines. Un motif est une suite de bases azotées ou d'acides aminés dans un ordre précis qui est retrouvée plusieurs fois dans un groupe de séquences (d'ADN ou protéiques). Les motifs MEME sont établis à partir de matrices de probabilité de position qui indiquent la probabilité de chaque lettre d'apparaître à une position du motif (les motifs MEME ne contiennent pas de lacunes). MEME prend en entrée un groupe de séquences d'ADN ou de séquences protéiques, et en sortie autant de motifs que demandé (Figure 38). Pour choisir automatiquement les meilleurs motifs, MEME utilise des techniques de modélisation statistique en comparant chaque motif à un modèle aléatoire : MEME construit tout d'abord une banque de séquences aléatoires où chaque position est indépendante, et où les bases sont choisies en fonction de leur fréquence dans la banque qu'on lui a soumise. MEME compare ensuite ces séquences aléatoires aux séquences fournies par l'utilisateur, et calcule la probabilité que chaque motif trouvé apparaisse par hasard, en donnant les valeurs suivantes :

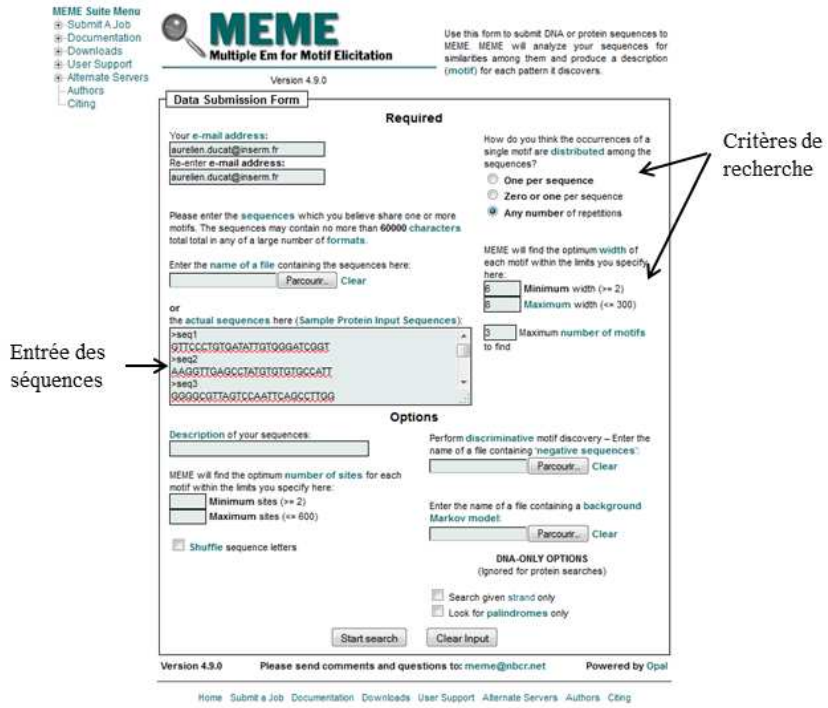
- le rapport de vraisemblance logarithmique (« log likelihood ratio ») : le rapport de vraisemblance logarithmique du motif est le logarithme du rapport de la probabilité des occurrences du motif au sein de l'ensemble des séquences données sur la probabilité des occurrences du motif au sein de l'ensemble des séquences aléatoires. Plus il est élevé, plus le motif trouvé est spécifique ;

- la E-value : il s'agit de la signification statistique du motif (plus la E-value est faible, plus le motif est statistiquement significatif). La E-value d'un motif est basé sur son rapport de vraisemblance logarithmique (« log likelihood ratio »), le nombre de bases constituant le motif, le nombre de séquences dans lesquelles il apparaît, les fréquences des bases dans l'ensemble des séquences données, et la taille de l'ensemble des séquences. La E-value est une estimation du nombre prévu de motifs ayant le même rapport de vraisemblance donnée (ou supérieur), la même largeur et le même nombre d'occurrences, que l'on pourrait trouver dans un modèle aléatoire. Il s'agit donc de la probabilité de trouver ce motif à cette fréquence dans un modèle aléatoire.

À partir de ces données, MEME représente les motifs trouvés sous forme d'un « LOGO » qui contient des piles de lettres à chaque position dans le motif. La hauteur totale de la pile est proportionnelle à la signification statistique de chaque lettre dans l'identification du motif (dans le détail, la hauteur de la pile est calculée à partir de l'entropie relative de cette position dans le motif en bits, autre valeur statistique dont je ne détaillerai pas le calcul ici). La hauteur des lettres individuelles d'une pile est la probabilité de la lettre à cette position multipliée par la hauteur totale de l'empilement.

Par ailleurs, MEME présente les occurrences (sites) du motif dans l'ensemble des séquences données, alignés les uns avec les autres. MEME présente également ces occurrences sous forme de diagramme. Chaque site est identifié par le nom de la séquence où il se produit, le volet (si les deux brins de séquences d'ADN sont utilisés), et la position dans la séquence où le site commence. Enfin, un schéma fonctionnel montrant un combiné de tous les motifs (non chevauchants) sur chacune des séquences contribuant aux motifs est présenté en fin d'analyse.

Figure 38 : Page d'entrée des données sur le logiciel MEME en ligne



Sur cette page, on entre nos séquences et on choisit nos critères de recherche : nombre d'occurrence du motif par séquence, taille minimale et maximale en paires de bases, nombre de motifs recherchés parmi les séquences données.

G. Gel retard

Des sondes simple-brin biotinylées et non-biotinylées contenant les motifs consensus identifiées au cours de la PCR-sélection ainsi que leur séquence complémentaire respective (biotinylées et non biotinylées) ont été commandées sur le site d'Eurogentec (http://www.eurogentec.com/eu-home.html) et sont représentées dans la Figure 39. Des sondes mutées non-biotinylées et leur séquence complémentaire ont également été commandées. Toutes ces sondes ont été hybridées avec leur brin complémentaire dans le milieu suivant (10 mM Tris, amené à pH8,0 avec HCl + 1 mM EDTA + 50 mM NaCl) en utilisant un programme thermique en gradient de température sur une machine PCR (95 °C pendant 5 minutes puis 70 cycles d'une minute allant de 95 °C à 25 °C en faisant 1 °C en moins à chaque cycle). L'efficacité de cette hybridation a été contrôlée sur gel d'agarose à 2 %. Ces sondes double brin ont été incubées 20 minutes avec les extraits protéiques des transfections, dans des proportions décrites dans le kit « LightShift® Chemiluminescent EMSA Kit » de chez Thermo Scientific. Un gel d'acrylamide à 6 % en condition non-dénaturante (voir composition en Annexe 7) a été coulé et ses sels préalablement éliminés par un « pre-run » à vide à 200 V pendant 30 à 60 minutes, effectué dans du TBE-0,5X froid. Les mélanges préparés ont été déposés sur ce gel, et la migration a été lancée à 100 V pendant une heure. Puis, un transfert de 30 minutes à 380 mA à 4 °C a été effectué sur une membrane de nylon HybondTM-N+ de chez GE Healthcare, préalablement hydratée pendant 10 minutes dans du TBE-0,5X. L'ADN transféré sur la membrane a ensuite été « cross-linké » par 2 expositions de 10 secondes aux UV (des liaisons covalentes avec la membrane de nylon ont été créées, étape indispensable pour éviter la perte des oligonucléotides lors des étapes ultérieures). La détection du signal a été réalisée grâce au kit « Chemiluminescent Nucleic Acid Detection Module » de chez Thermo Scientific qui utilise le système streptavidine-peroxydase de raifort + luminol. La révélation a été réalisée sur un film photographique.

Figure 39 : Séquences des sondes utilisées pour le retard sur gel

① Sonde biotinylée (●) n° 1 avec 3 sites CATYTCACGG (consensus n° 1) : 3xSTRE1-biot
●-AGAGC CATYTCACGG AGAGC CATYTCACGG AGAGC CATYTCACGG AGAGC-●

② Sonde biotinylée (●) n° 2 avec 3 sites GGTGYGGAMA (consensus n° 2) : 3xSTRE2-biot
●-GCTAT GGTGYGGAMA GCTAT GGTGYGGAMA GCTAT GGTGYGGAMA GCTAT-●

③ Sonde non-biotinylée compétitrice n° 1 (contenant 3 sites consensus n° 1) : 3xSTRE1
AGAGC CATYTCACGG AGAGC CATYTCACGG AGAGC CATYTCACGG AGAGC

④ Sonde non-biotinylée compétitrice n° 2 (contenant 3 sites consensus n° 2) : 3xSTRE2
GCTAT GGTGYGGAMA GCTAT GGTGYGGAMA GCTAT GGTGYGGAMA GCTAT

⑤ Sonde non-biotinylée mutée n° 1 (contenant 3 sites consensus n° 1 mutés) : 3xmSTRE1
AGAGC **CACYTCATGG** AGAGC **CACYTCATGG** AGAGC **CACYTCATGG** AGAGC

⑥ Sonde non-biotinylée mutée n° 2 (contenant 3 sites consensus n° 2 mutés) : 3xmSTRE2
GCTAT **GATGYTGAMA** GCTAT **GATGYTGAMA** GCTAT **GATGYTGAMA** GCTAT

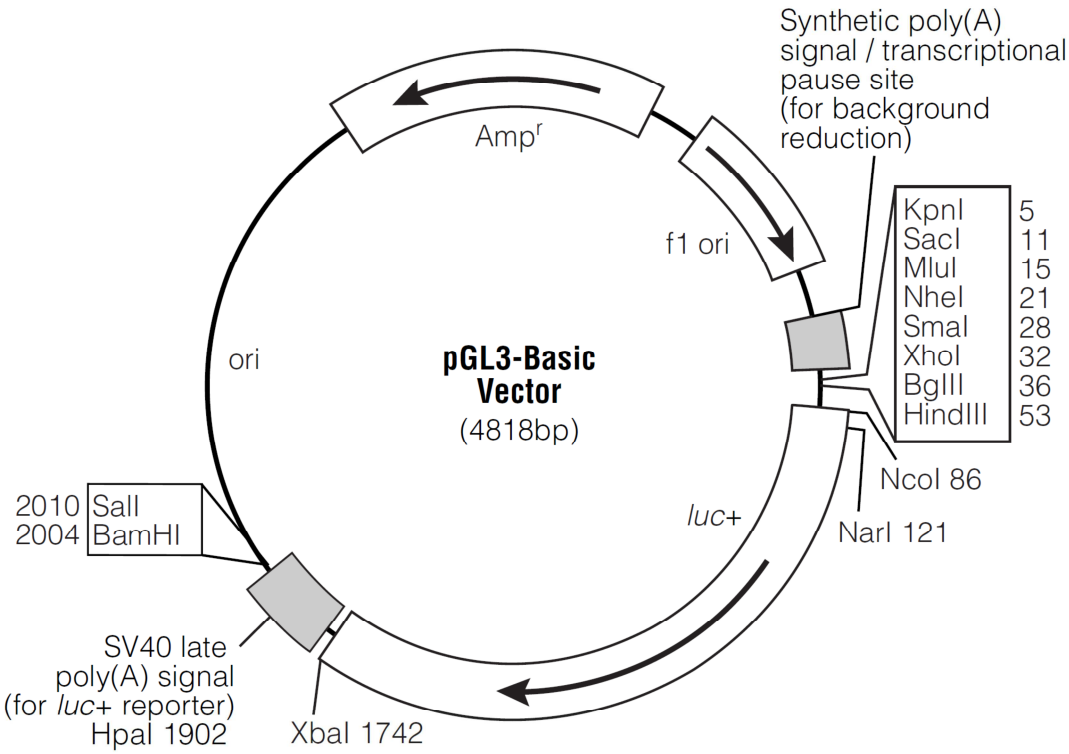
Des sondes simple-brin biotinylées et non-biotinylées contenant les motifs consensus identifiées au cours de la PCR-sélection ainsi que leur séquence complémentaire respective (biotinylées et non biotinylées) ont été commandées sur le site d'Eurogentec (http://www.eurogentec.com/eu-home.html). Des sondes mutées non-biotinylées et leur séquence complémentaire ont également été commandées. Les quelques bases séparant 2 motifs consensus sur une même séquence ont été choisies grâce à un tableau d'occurrences qui sera détaillé plus loin (les bases azotées les moins fréquentes ont été choisies).

H. Essai luciférase

Pour valider l'effet fonctionnel des séquences consensus trouvées (censées être de possibles sites de fixation de STOX1), j'ai intégré ces séquences dans un vecteur contenant l'ADNc de la luciférase firefly déjà disponible au laboratoire, dérivé du pGL3-Basic de Promega (Madison, États-Unis), dans lequel un CMV minimal a été intégré entre les sites des enzymes de restrictions XhoI et HindIII (Figure 40). Les séquences intégrées correspondent aux sondes hybridées qui ont été utilisées pour le retard sur gel. Pour une intégration en bouts francs, ces séquences double-brin doivent être phosphorylées : cette phosphorylation a été réalisée avec l'enzyme T4 Polynucléotide kinase de chez Thermo Scientific. Le vecteur a ensuite été digéré avec l'enzyme de restriction SmaI de chez Life Technologies™, puis déphosphorylé avec de la SAP (*Shrimp Alkaline Phosphatase*) de chez Affymetrix® et enfin purifié sur colonne avec le kit « NucleoSpin® Gel and PCR Clean-up » (Macherey-Nagel). La ligation a été réalisée avec de la ligase T4 de chez Thermo Scientific pendant toute une nuit en présence du vecteur digéré purifié et des inserts phosphorylés purifiés présents en excès molaire (5 à 10 fois). La transformation de bactéries, la sélection de colonies intéressantes, l'extraction d'ADN plasmidique et le séquençage ont eu lieu dans les mêmes conditions que pour la construction des vecteurs d'expression Flag-STOX1. La PCR sur clones a été réalisée avec des amorces universelles situées de part et d'autre des inserts attendus (GL2 et RV3). Le principe de l'essai luciférase a ensuite été de faire plusieurs transfections de plusieurs vecteurs dans des cellules COS avec de la lipofectamine, comme décrit précédemment (II.B). Plusieurs transfections ont été réalisées dans des plaques 24 puits avec différentes combinaisons de vecteurs : le vecteur renilla (5 ng, qui a servi à normaliser les résultats sur l'efficacité relative de la transfection), le

vecteur pGL3 avec nos inserts (270 ng), le vecteur inducteur pCMX-STOX1A (200 ng). Chaque transfection a été réalisée en « sixplicat ». Les quantités de milieu, de plasmide et de lipofectamine dans ces boîtes ont été calculées au *prorata* de la surface cellulaire (un puits d'une plaque 24 puits faisant 1,862 cm²). La révélation a été réalisée en utilisant le kit « Dual-Luciferase® Reporter Assay System » de chez Promega. Les analyses statistiques ont été réalisées avec le logiciel Excel®2010. Le test Mann-Whitney a été utilisé avec un seuil de signification statistique à 0,05.

Figure 40 : Carte du vecteur pGL3-Basic (Promega)



D'après le manuel technique du « pGL3 Luciferase Reporter Vectors » disponible sur le site de Promega (http://france.promega.com/). Ce vecteur contient l'ADNc de la luciférase *firefly* et de multiples sites d'enzymes de restriction en amont, ce qui nous permet d'insérer des séquences en amont de ce gène et de tester leur effet sur l'expression de la luciférase.
luc+ : ADNc codant pour la luciférase *firefly* modifiée.
Amp^r : gène conférant une résistance à l'ampicilline chez *Escherichia coli*.
f1 ori : origine de réplication provenant de phage filamenteux
ori : origine de réplication dans *Escherichia coli*.
Les flèches dans *luc*+ et le gène Amp^r indiquent le sens de la transcription. La flèche dans le f1 ori indique la direction de la synthèse du brin ADN simple brin.

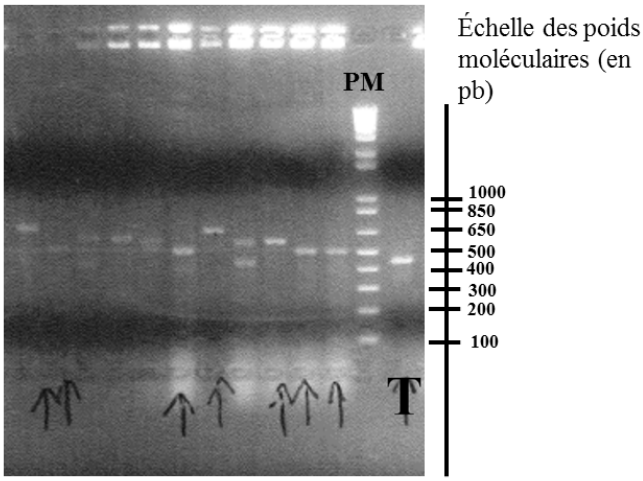
III. Résultats

A. Construction du vecteur plasmidique et validation en Western blot

J'ai réalisé des constructions plasmidiques permettant l'expression de protéines chimériques contenant un peptide spécifique appelé « Flag » (DYKDDDDK) : Flag-STOX1A et Flag-STOX1B. Pour cela, j'ai réalisé un clonage en bout cohésifs d'un insert contenant la séquence de 2 Flag (Figure 34). L'intégration de cet insert ne permet l'expression de la protéine chimérique que si l'insert s'intègre dans un sens adéquat (pour avoir le bon cadre de lecture et les bons codons). Afin d'obtenir plusieurs type de protéines chimériques avec un nombre de Flag différents, j'ai mis l'insert en fort excès molaire (10 fois) lors de la ligation, favorisant ainsi la ligation de plusieurs inserts en tandem dans le plasmide.

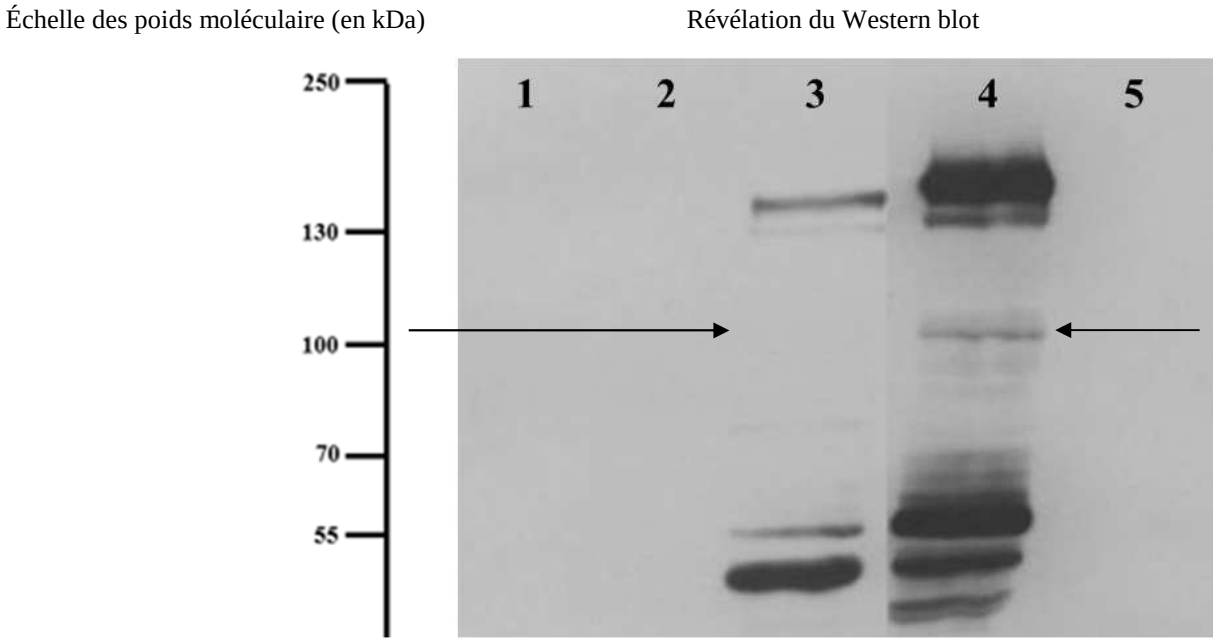
Pour Flag-STOX1A, j'ai obtenu 7 clones intéressants ayant intégré au moins 1 insert (Figure 41). Après séquençage, il s'est avéré qu'un plasmide avait intégré un insert dans le bon sens (donc 2 séquences Flag, voir Figure 34), qu'un autre plasmide avait intégré 2 inserts dans le bon sens (4 séquences Flag) et qu'un autre plasmide avait intégré 3 inserts dans le bon sens (6 séquences Flag). À la suite de cela, j'ai réalisé plusieurs transfections suivies de Western blot avec un anti-Flag afin de vérifier l'expression de ma protéine. J'ai pu remarquer au fur et à mesure de mes transfections et de mes WB, que les protéines chimériques contenant le plus de séquences Flag engendraient un signal plus fort en WB (et ce, de manière reproductible). C'est pourquoi je me suis concentré sur la protéine chimérique 6-Flag-STOX1A (dont la séquence complète figure en Annexe 3). En revanche, j'obtenais toujours plusieurs bandes à différents poids moléculaires en WB suite à mes transfections avec ce plasmide (Figure 42). La plupart des bandes se trouvant en dessous du poids moléculaire attendu (environ 114 kDa), je me suis ainsi demandé s'il ne s'agissait pas de produit de dégradation, et j'ai donc essayé de réaliser des transfections en présence de MG132, un inhibiteur du protéasome. J'ai pu vraiment observer une différence dans la taille et le nombre de bandes en WB (Figure 42), dont une au bon poids moléculaire (indiquée par une flèche sur la Figure 42) : il y avait une diminution du signal pour la bande inférieure à 55 kDa, une augmentation pour les autres bandes de plus grand poids moléculaire, et apparition d'autres bandes à divers poids moléculaires non attendus. Les signaux de poids moléculaire inférieur à celui attendu pouvaient correspondre très certainement à des formes tronquées ou modifiées de la protéine car il n'y avait aucun signal pour les contrôles négatifs. En ce qui concerne les bandes de plus haut poids moléculaire en revanche, nous avons pu penser à une forme modifiée de la protéine : ubiquitinylée (ou plutôt poly-ubiquitinylée à ce moment-là car une molécule d'ubiquitine ne fait que 8,5 kDa), phosphorylée (idem plutôt poly-phosphorylée).

Figure 41 : Électrophorèse d'une PCR sur clones pour la construction Flag-STOX1A



Électrophorèse des produits de PCR sur les clones issus de la transformation de bactéries avec des constructions Flag-STOX1A. PCR réalisée avec les primers T7P et 360R (voir Figure 34). Migration dans un gel d'agarose à 2 % + BET (bromure d'éthidium) au 20 000ème. Taille attendue en l'absence d'insert : 430 pb ; taille attendue en présence d'un insert : 500 pb. PM : poids moléculaire (1 Kb Plus DNA Ladder, Life Technologies). T : témoin sans insert. Les flèches désignent les clones sélectionnés.

Figure 42 : Western blot de validation pour Flag-STOX1A

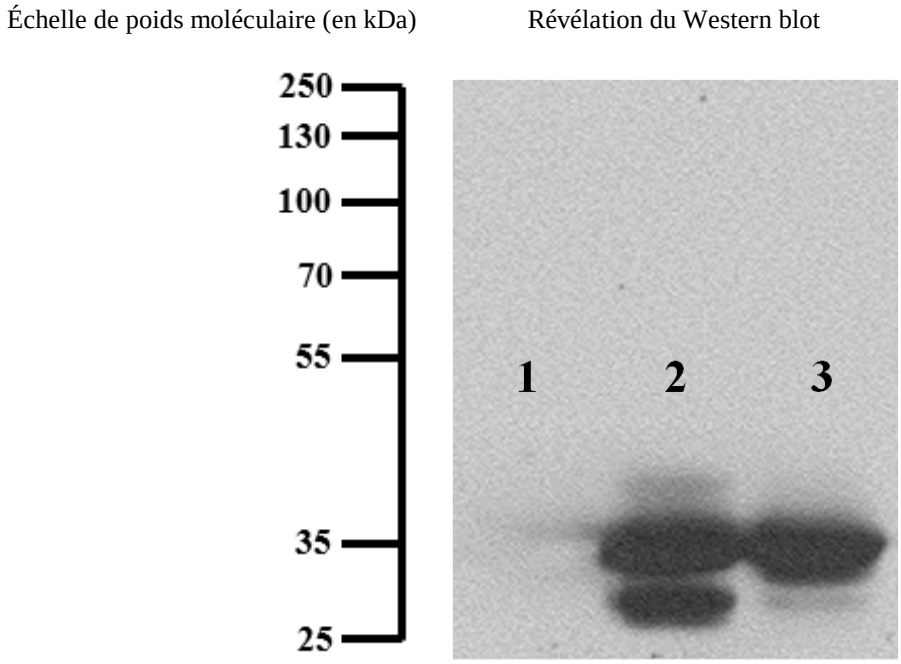


Migration dans un gel à 7 %. Anticorps primaire anti-Flag M2 (Sigma F3165) dilué au 2 500ème, révélation avec le kit « ECL SuperSignal® West Pico Chemiluminescent Substrate » de chez Thermo Scientific pendant 3 secondes. Référence de l'échelle de poids moléculaires utilisée : « PageRuler^{TM} Plus Prestained Protein Ladder » de chez Thermo Scientific. La flèche désigne la taille attendue de la protéine 6Flag-STOX1A.
Description des puits :
1 : Transfection vide
2 : Transfection avec le plasmide pCMX-STOX1A
3 : Transfection avec le plasmide pCMX-6Flag-STOX1A
4 : Transfection avec le plasmide pCMX-6Flag-STOX1A + ajout de MG132
5 : Transfection avec le plasmide pCMX-STOX1A + ajout de MG132

En parallèle, j'ai également construit des plasmides Flag-STOX1B : STOX1B étant une isoforme plus petite et contenant comme STOX1A le domaine présumé de fixation à l'ADN, j'ai supposé qu'elle serait plus stable lors des transfections et moins sujette à de multiples dégradations protéiques intracellulaires. Lors de mes constructions, j'ai obtenu 5 clones intéressants ayant intégré au moins 1 insert. Après séquençage, il s'est avéré qu'un plasmide avait intégré un insert dans le bon sens (donc 2 séquences Flag, voir Figure 5), et qu'un autre plasmide avait intégré 2 inserts dans le bon sens (4 séquences Flag). J'ai fini par sélectionner ce dernier car le signal obtenu après WB était bien plus fort et ce, de manière reproductible (la séquence protéique de 4Flag-STOX1B figurent en Annexe 5). En effet, plusieurs transfections suivies de WB m'ont permis de mettre en évidence la présence de cette protéine chimérique dans nos extraits par un signal fortement spécifique à la taille attendue (30 kDa).

Au vu de tous ces résultats, j'ai donc pris la décision de réaliser la PCR-sélection en utilisant la protéine chimérique 4Flag-STOX1B, qui semblait stable en transfection dans des cellules COS-7, et dont l'expression était vérifiable en WB. J'ai donc réalisé une transfection à plus grande échelle dans des boîtes rondes de culture cellulaire de 60,1 cm² (TPP®). Le WB de validation de cette transfection est présenté en Figure 43 : on remarque l'apparition de deux bandes, une à la bonne taille et une légèrement plus basse qui pourrait éventuellement correspondre à une forme tronquée de 4Flag-STOX1B.

Figure 43 : Western blot de validation pour Flag-STOX1B



Anticorps primaire anti-Flag M2 (Sigma F3165) dilué au 2 500ème, révélation avec le kit « ECL SuperSignal® West Pico Chemiluminescent Substrate » de chez Thermo Scientific pendant 3 secondes. Référence de l'échelle de poids moléculaire utilisée : « PageRuler™ Plus Prestained Protein Ladder » de chez Thermo Scientific.
Description des puits :
1 : Transfection vide
2 : Transfection pour la PCR-sélection
3 : Témoin positif (transfection test)

B. Résultat de la PCR-sélection : obtention de séquences

Le principe de notre PCR-sélection va être de purifier des oligonucléotides sur lesquels un facteur transcriptionnel se fixe, au sein d'une banque d'oligonucléotides aléatoires. Le facteur transcriptionnel va être ici la protéine chimérique 4Flag-STOX1B. Des extraits protéiques issus de cellules transfectées avec le vecteur pCMX-4Flag-STOX1B (donc contenant la protéine chimérique 4Flag-STOX1B) et des oligonucléotides aléatoires ont été mis à incuber, puis le mélange a été immunoprécipité sur des billes couplées à des anticorps anti-Flag (reconnaissant seulement le Flag). Une élution finale nous a permis d'obtenir les séquences d'ADN retenues, et une PCR nous a permis de les amplifier. Cinq étapes consécutives ont été réalisées. Les oligonucléotides sélectionnés par cette technique ont ensuite été clonés puis séquencés (la Figure 35 résume le tout).

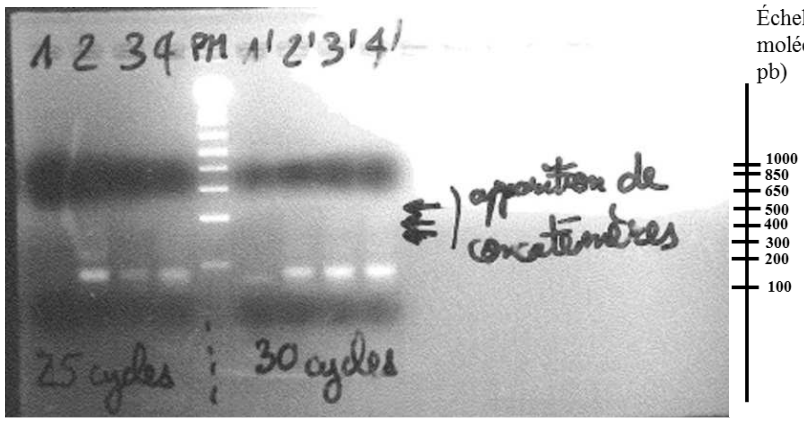
En parallèle de cette PCR-sélection, j'ai effectué une autre PCR-sélection sur le même principe mais en utilisant cette fois-ci des extraits protéiques issus de cellules transfectées avec le vecteur pCMX-STOX1B (sans Flag) : cette expérience a constitué mon contrôle négatif.

Au cours des étapes successives de purification, je me suis aperçu en faisant migrer nos produits de PCR sur gel qu'il y avait apparition de concatémères (Figure 44). Afin de ne pas sélectionner ceux-ci au cours de l'opération (car ils auraient pu être sélectionnés par la présence d'un seul site de fixation pour un grand fragment, alors qu'il est préférable d'avoir un site sélectionné par petit fragment de 76 pb), j'ai réalisé à la fin de chaque tour deux PCR différentes, avec un nombre de cycles différent. Les produits que je sélectionnais pour réaliser le tour de sélection suivant étaient ceux qui présentaient le moins de concatémères. On remarquera aisément qu'en l'absence de protéine « flaguée », la PCR a fonctionné malgré tout, ce qui nous a indiqué qu'il y avait toujours des oligonucléotides après les rinçages, même s'il n'y avait pas de protéines « flaguées » pour les retenir sur les billes. Cela passait donc soit par l'intermédiaire de protéines aspécifiques, soit via l'anticorps, soit directement sur les billes.

Afin de séquencer chaque fragment contenant normalement un site de fixation, j'ai inséré les produits de PCR du 5ème tour de sélection dans un vecteur TOPO-TA. Cela m'a permis d'obtenir 58 séquences issues de la PCR-sélection avec la protéine 4Flag-STOX1B, et 44 séquences issues du contrôle négatif de la PCR-sélection réalisée avec la protéine STOX1B (voir Annexe 8).

Figure 44 : Exemple d'électrophorèse de produits de PCR au cours d'un tour de PCR-sélection

Électrophorèse des produits de PCR au cours de la PCR-sélection. PCR réalisée avec les amorces sens et anti-sens (voir Figure 35). Migration dans un gel d'agarose à 2 % + BET au 20 000ème. Taille attendue : 76 pb. PM : poids moléculaire (1 Kb Plus DNA Ladder, Life Technologies). Les flèches désignent l'apparition de concatémères.



Légende :
1 et 1' : témoins négatifs (bande aspécifique plus basse que les autres)
2 et 2' : témoins positifs
3 et 3' : PCR sur oligonucléotides sélectionnés avec 4Flag-STOX1B
4 et 4' : PCR sur oligonucléotides sélectionnés avec STOX1B

C. Logiciel MEME : deux séquences consensus ont été identifiées

1. Résultats pour 4Flag-STOX1B

L'analyse des résultats a été faite sur les 58 séquences obtenues (voir Annexe 8), avec les réglages suivants :
- distribution des occurrences de motifs : n'importe quel nombre de répétitions.
- nombre de différents motifs : 3
- largeur minimale de motif : 6
- largeur maximale de motif : 8

Pour déterminer la taille du motif attendu, je me suis basé sur les données déjà publiées sur les motifs obtenus des facteurs de transcription FOX (Benayoun *et al.*, 2008), qui décrivaient un motif de 8 bases.

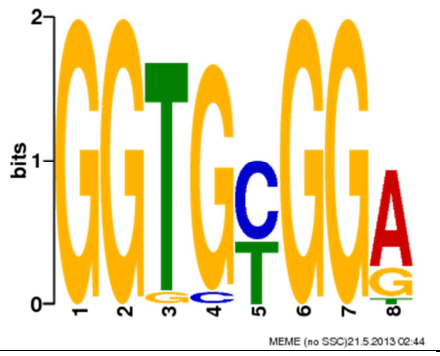
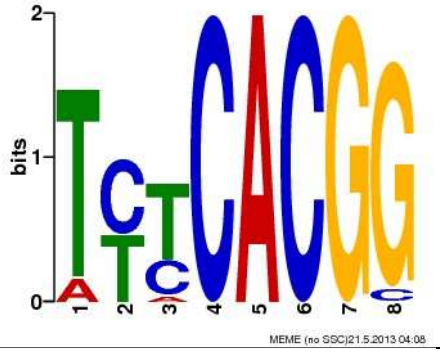
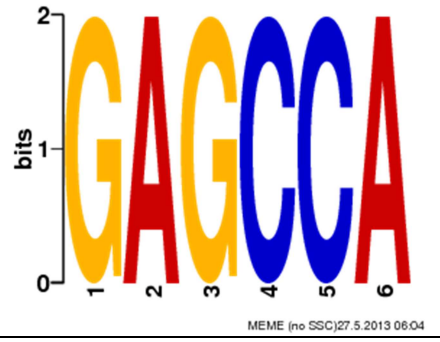
Les résultats de la soumission de ces séquences sur MEME sont présentés en Figure 45 : le logiciel a mis en évidence deux séquences consensus significatives avec des E-value de $1{,}5.10^{-11}$ pour le motif 1 et de $4{,}1.10^{-9}$ pour le motif 2 (Figure 45).

La suite de l'analyse a consisté à aligner manuellement toutes nos séquences sur Excel en repérant les bases du/des motif(s) (autour des lettres les plus occurrentes), et à construire un tableau d'occurrences pour chaque motif consensus trouvé (Figure 46), qui donnait le pourcentage de chaque base pour chaque position. Par rapport à l'analyse par MEME, cette analyse permet d'inclure les séquences tronquées, les motifs qui se répètent plusieurs fois dans une séquence, les motifs un peu plus longs, ce qui a permis de trouver le motif 1 dans 23 séquences (alors que MEME ne le trouvait que dans 18).

Une donnée tout à fait intéressante à analyser a été le diagramme combiné : il montrait de plus que les deux séquences se retrouvaient souvent (pour 13 séquences sur 16) sur les mêmes séquences (Figure 47).

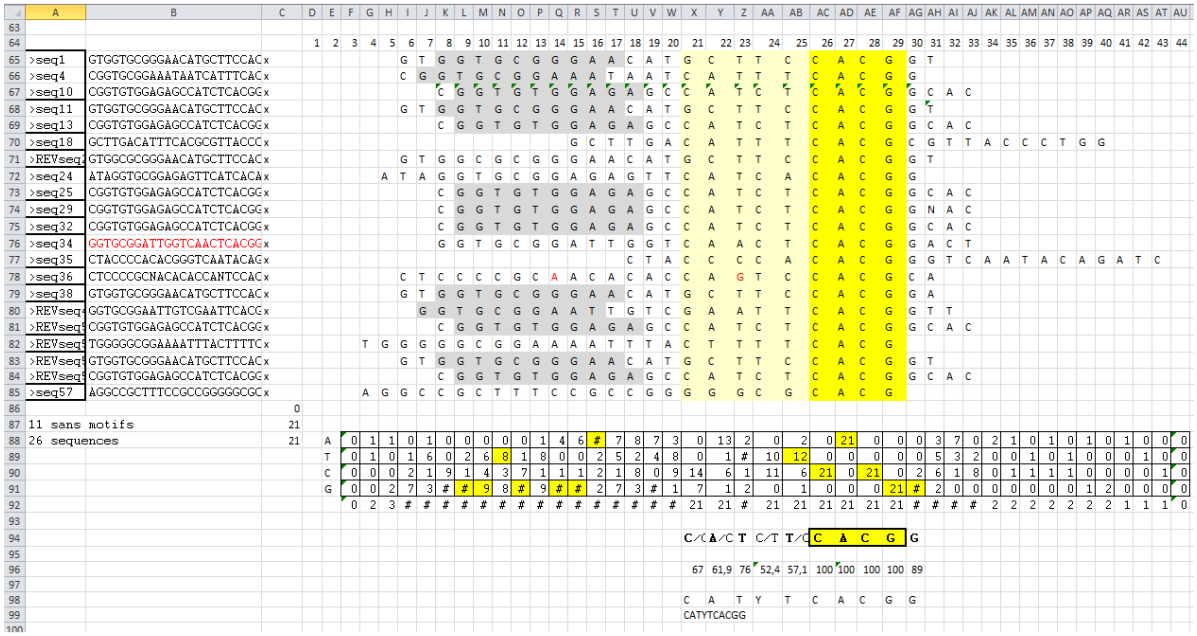
Au final, ces tableaux m'ont donc permis de trouver 23 fois le motif n^{o} 1 dans 21 séquences et de trouver 21 fois le motif n^{o} 2 dans 21 séquences sur 58. Cette analyse m'a permis enfin de faire ressortir deux séquences consensus de 10 bases (Figure 48).

Figure 45 : Séquences obtenues avec le logiciel MEME à partir des séquences issues de la PCR-sélection réalisée sur la protéine 4Flag-STOX1B

Motif 1 : - E-value : $1{,}5.10^{-11}$ - nombre de sites : 18		➔ a été appelé par la suite séquence consensus n° 2
Motif 2 : - E-value : $4{,}1.10^{-9}$ - nombre de sites : 17		➔ a été appelé par la suite séquence consensus n° 1
Motif 3 : - E-value : $1{,}8.10^{2}$ - nombre de sites : 9		

Pour chaque motif trouvé, MEME donne la E-value. Les lettres dans les logos sont toujours colorées de la même façon (Adénine en rouge, Cytosine en bleu, Guanine en orange, Thymine en vert).

Figure 46 : Alignement des séquences sur Excel (exemple sur le motif 2)



Toutes nos séquences ont été alignées manuellement sur Excel en repérant les bases du/des motif(s) (autour des lettres les plus occurrentes). Par rapport à l'analyse par MEME, cette analyse permet d'inclure les séquences tronquées, les motifs qui se répètent plusieurs fois dans une séquence, les motifs un peu plus longs.

Figure 47 : Diagramme combiné des séquences consensus n° 1 et 2

Lorsque le brin d'ADN est spécifié, « + » désigne la séquence soumise, et « - » signifie le brin complémentaire de cette séquence.

Figure 48 : Tableaux des occurrences après alignement des séquences

Un tableau d'occurrences a été construit pour chaque motif consensus trouvé : il donne le pourcentage de chaque base pour chaque position.

Motif 1 (aligné sur GGNNNGG) :

A	0	0	3	3	0	0	0	13	9	16
T	0	0	17	1	12	0	0	3	3	4
C	0	0	2	1	11	0	0	1	1	2
G	23	23	1	18	0	23	23	6	10	1
Total	23	23	23	23	23	23	23	23	23	23
Consensus	**G**	**G**	**T**	**G**	**T**/C	**G**	**G**	**A**	A/**C**	**A**
Pourcentage du plus occurrent	100	100	74	78	52	100	100	57	43	70

Consensus n° 2 : GGTG[T/C]GGA[A/C]A ou GGTGYGGAMA

Motif 2 (aligné sur CACG) :

A	0	13	2	0	2	0	21	0	0	0
T	0	1	16	10	12	0	0	0	0	0
C	14	6	1	11	6	21	0	21	0	2
G	7	1	2	0	1	0	0	0	21	17
Total	21	21	21	21	21	21	21	21	21	19
Consensus	**C**	**A**	**T**	T/**C**	**T**	**C**	**A**	**C**	**G**	**G**
Pourcentage du plus occurrent	66,7	61,9	76	52,4	57,1	100	100	100	100	89

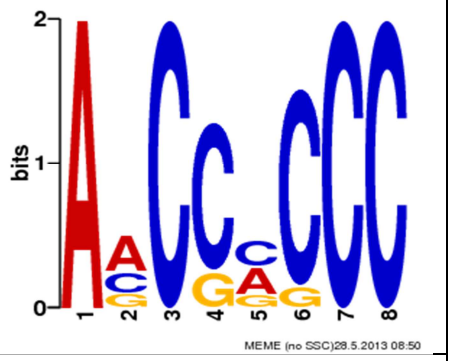
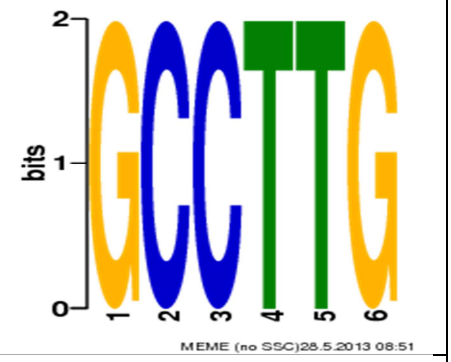
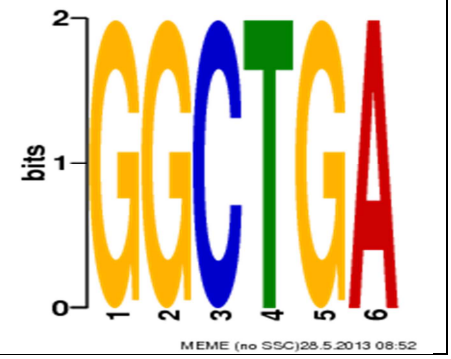
Consensus n° 1 : CAT[T/C]TCACGG ou CATYTCACGG

2. Résultats pour STOX1B (contrôle négatif)

L'analyse des résultats a été faite sur les 44 séquences obtenues, en utilisant les mêmes réglages que pour Flag-STOX1B :
- distribution des occurrences de motifs : n'importe quel nombre de répétitions.
- nombre de différents motifs : 3
- largeur minimale de motif : 6
- largeur maximale de motif : 8

Aucune séquence significative n'a été trouvée par MEME (Figure 49). Cela confirmait qu'en l'absence de protéine « flaguée », les oligonucléotides n'étaient pas sélectionnés sur un motif particulier.

Figure 49 : Séquences obtenues avec le logiciel MEME à partir des séquences issues de la PCR-sélection réalisée sur la protéine STOX1B.

Motif 1 : - E-value : $3{,}4.10^{+2}$ - nombre de sites : 10	
Motif 2 : - E-value : $5{,}2.10^{+4}$ - nombre de sites : 2	
Motif 3 : - E-value : $5{,}2.10^{+4}$ - nombre de sites : 2	

Noter l'absence de E-value faible (pas de motifs significativement enrichis).

3. Bilan

La PCR-sélection réalisée avec la protéine 4Flag-STOX1B a fait ressortir 2 séquences consensus possibles de 10 paires de bases. Ces séquences vont être appelées dans la suite de ce rapport STRE1 et STRE2 (pour *STOX1 Reponsive Element*).

STRE1 : CATYTCACGG
STRE2 : GGTGYGGAMA

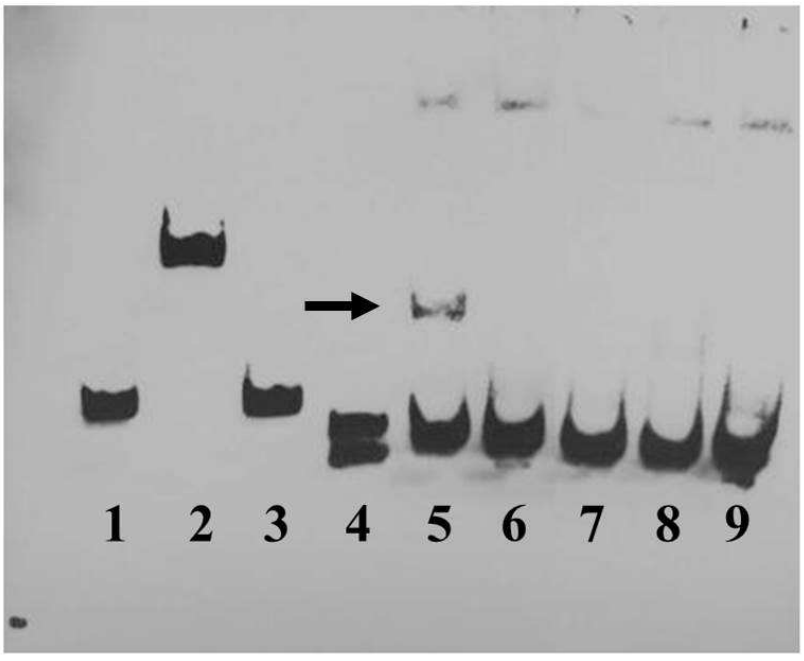
D. Validation en retard sur gel

L'expérience de retard sur gel a permis de valider l'interaction physique entre la protéine et les séquences consensus trouvées.

J'ai donc commandé des sondes biotinylées et non-biotinylées contenant soit 3 sites STRE1 (sonde 3xSTRE1), soit 3 sites STRE2 (sonde 3xSTRE2). Pour la mise au point des conditions, j'ai commencé par préparer 3 mélanges pour chaque sonde : sonde biotinylée seule ; sonde biotinylée + extraits protéiques ; sonde biotinylée + extraits protéiques + la même sonde non biotinylées en excès molaire (200 fois).

Dans cette première expérience, les sondes utilisées étaient : la sonde fournies dans le kit servant de témoin (appartenant au système EBNA : *Epstein-Barr Nuclear Antigen*), la sonde 3xSTRE1 et la sonde 3xSTRE2. J'ai ensuite déposé sur gel ces différents mélanges. Cette expérience a été réalisée deux fois indépendamment et a donné le même résultat (Figure 50). On a observé dans un premier temps que la séquence consensus n° 2 ne semblait pas créer de retard. Nous nous sommes donc concentrés sur la séquence n° 1, qui nous donnait déjà des résultats préliminaires intéressants. En effet, il y avait apparition d'un retard lorsque l'on incubait la sonde biotinylée avec des extraits protéiques, retard qui disparaissait lorsqu'on rajoutait cette même sonde non-biotinylée en excès molaire (expérience de compétition).

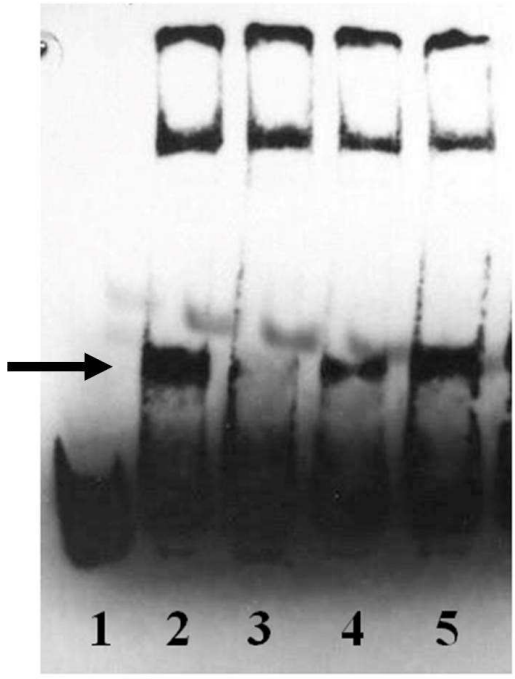
Figure 50 : Première expérience de retard sur gel



Description des puits :
1 : ADN biotinylé contrôle EBNA
2 : ADN biotinylé contrôle EBNA + extraits protéiques EBNA
3 : ADN biotinylé contrôle EBNA + extraits protéiques EBNA + ADN non-biotinylé contrôle EBNA en excès molaire (200 fois)
4 : STRE1biot
5 : STRE1biot + extraits protéiques (contenant 4Flag-STOX1B)
6 : STRE1biot + extraits protéiques (contenant 4Flag-STOX1B) + STRE1
7 : STRE2biot
8 : STRE2biot + extraits protéiques (contenant 4Flag-STOX1B)
9 : STRE2biot + extraits protéiques (contenant 4Flag-STOX1B) + STRE2
La flèche indique le retard observé avec la protéine 4Flag-STOX1B.

J'ai ensuite réalisé une autre expérience afin de tester si la bande retard était bien spécifique de notre protéine d'intérêt. Ainsi, j'ai préparé plusieurs mélanges utilisant la séquence consensus n° 1 : un mélange contenant uniquement de la sonde biotinylée, un mélange contenant la sonde biotinylée et des extraits protéiques, un mélange contenant la sonde biotinylée avec des extraits protéiques et la même sonde non-biotinylée en excès molaire (200 fois), un mélange contenant la sonde biotinylée avec des extraits protéiques et une sonde mutée non-biotinylée en excès molaire (200 fois), et un mélange contenant la sonde biotinylée avec des extraits protéiques et l'anticorps anti-Flag (Figure 51). Pour ce dernier puits, si la séquence que nous avons trouvé était spécifique de notre protéine, nous aurions dû obtenir un « supershift » (c'est-à-dire une bande plus haute que celle apparue avec la sonde biotinylée et les extraits protéiques seuls, en raison de la formation de complexes ADN-facteur transcriptionnel-anticorps). L'absence de « supershift » a suggéré que ce n'était pas STOX1 qui se liait directement à la séquence d'ADN ; en revanche, la bande retard qui apparaissait sur tous nos gels semblait être spécifique de la séquence STRE1, car elle disparaissait lorsqu'on ajoutait la sonde non-biotinylée en excès molaire et elle réapparaissait lorsqu'on ajoutait la sonde mutée non-biotinylée en excès molaire.

Figure 51 : Retard sur gel sur la séquence consensus n° 1



Description des puits :
1 : STRE1biot
2 : STRE1biot + 4Flag-STOX1B
3 : STRE1biot + 4Flag-STOX1B + STRE1
4 : STRE1biot + 4Flag-STOX1B + mSTRE1
5 : STRE1biot + 4Flag-STOX1B + Anticorps anti-Flag M2 (Sigma F3165)
La flèche indique le retard observé avec la protéine 4Flag-STOX1B.

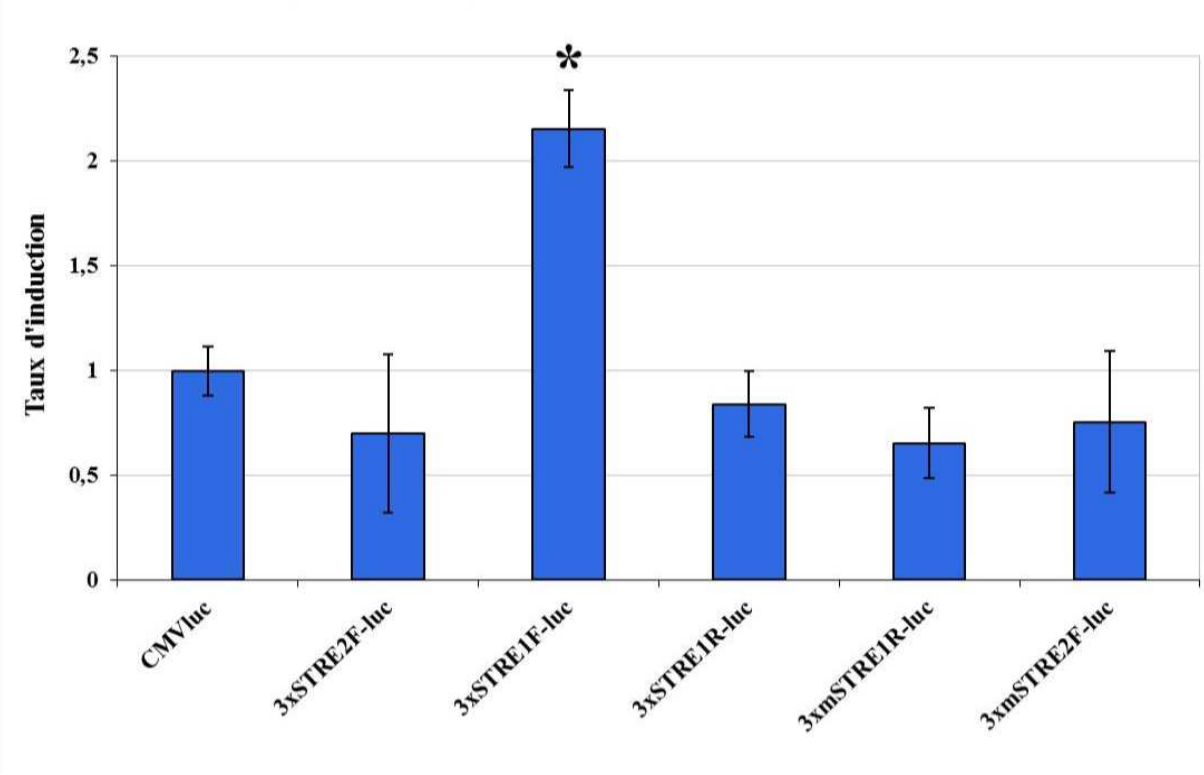
Ces premières expériences de validation en gel retard nous ont permis de dire que :

- la séquence consensus n° 2 n'était pas fixée par une protéine contenue dans ces extraits nucléaires dans les conditions utilisées pour ce retard sur gel.
- La séquence consensus n° 1 liait spécifiquement un facteur contenu dans les extraits protéiques. *A priori*, ce facteur n'était pas STOX1.

E. Premier essai luciférase

Je n'ai pas encore réussi à obtenir tous les plasmides recombinants : j'ai pu intégrer (« F » désigne une intégration dans le bon sens, « R » dans le sens inverse) : STRE1-F, STRE1-R, mSTRE1-R, STRE2-F et mSTRE2-F. J'ai pu réaliser un premier essai luciférase avec le vecteur inducteur pCMX-STOX1A (en effet, même si la PCR-sélection a été faite avec STOX1B, c'est l'isoforme A qui contient un domaine de transactivation et est censé avoir un effet biologique), dont les résultats sont présentés en Figure 52. Ces résultats montraient que le motif n° 1 semblait induire une activité plus importante de la luciférase (2 fois), ce qui concordait avec les résultats du retard sur gel.

Figure 52 : Premier essai luciférase



Les histogrammes représentent le taux d'induction de l'activité luciférase, c'est-à-dire le rapport entre l'activité luciférase en présence de STOX1A (expression via le plasmide pCMX-STOX1A) sur l'activité sans STOX1A (transfection d'un vecteur pCMX vide). * : Test Mann-Whitney orienté significatif ($p < 0,05$).

IV. Discussion

STOX1 a été clairement montré comme un gène important dans la physiopathologie de la prééclampsie. En effet, *in vitro*, sa surexpression dans un modèle cellulaire de trophoblaste (les cellules de choriocarcinome JEG-3) mime une partie des effets de la prééclampsie (corrélation des altérations transcriptomiques). De plus, *STOX1* a été montré surexprimé dans des biopsies de premier trimestre de femmes qui ont développé une prééclampsie par la suite. Enfin, des études *in vivo* ont montré que la surexpression de ce gène chez la souris reproduit un phénotype prééclamptique. En revanche, l'étude du rôle précis de *STOX1*, aussi bien à l'échelle cellulaire qu'à l'échelle de l'organisme est balbutiante. Il apparait donc aujourd'hui important d'investir un effort de recherche sur ce gène pour aider à avancer sur la compréhension globale de la physiopathologie de la maladie, et envisager des thérapeutiques. De plus, il s'agit d'un gène exprimé de façon ubiquitaire, donc une meilleure connaissance de sa(ses) fonction(s) pourrait également avoir un impact sur l'étude d'autres maladies.

Mon travail au cours de mon stage de recherche a consisté à identifier le site de fixation à l'ADN de la protéine STOX1. Le but ultérieur de cette opération était de pouvoir prédire les gènes que STOX1 régule de façon directe en tant que facteur de transcription, en croisant nos données avec les séquences promotrices de gènes d'intérêt. Les résultats de PCR-sélection obtenus ont mis en évidence deux séquences possibles pour le site de fixation de STOX1. Une de ces séquences semble intéressante à investiguer au vu des premiers résultats encourageants de validation en retard sur gel et essai luciférase. Cependant, ces résultats préliminaires ne sont pas suffisants et les recherches méritent d'être poursuivies. Toutefois, je n'ai pas obtenu d'évidence que les séquences trouvées étaient des sites de fixation directs de STOX1. À partir de là, plusieurs hypothèses peuvent être envisagées.

Tout d'abord, la sélection spécifique de nos oligonucléotides peut ne pas avoir fonctionné : en effet, il est possible que la séquence Flag rajoutée en N-terminal ait gêné le repliement correct de la protéine ou encombré le site de fixation à l'ADN. De plus, nous avions observé, lors des validations en Western blot, une bande de dégradation : cette protéine dégradée pouvait possiblement interférer avec les expériences de « supershift » en générant une forme sans Flag capable d'interagir avec la séquence cible sans être reconnue par l'anticorps.

Si l'expérience de sélection a fonctionné, il est tout à fait possible d'avoir mis en évidence un partenaire de STOX1 interagissant directement avec la séquence et aussi avec STOX1. C'est une hypothèse qui vient assez vite à l'esprit quand on regarde la co-sélection de deux séquences consensus (Figure 47) : s'agit-il de deux partenaires de STOX1, chacun étant sur un site différent ? Ceci pose des questions sur le mécanisme d'action de STOX1 : comment STOX1 agit-il sur la transcription ? A-t-il besoin de co-facteurs ou de partenaires ? Interagit-il avec d'autres facteurs transcriptionnels ? Se lie-t-il directement à l'ADN ? Finalement aujourd'hui, quels sont les arguments pour dire que STOX1 est un facteur de transcription ? Nous avons l'analyse bio-informatique nous prédisant un domaine de liaison acide, mais nous avons aussi les modifications transcriptomiques suite à sa surexpression dans un modèle cellulaire. Même si de fortes preuves existent donc pour dire que STOX1 est un facteur transcriptionnel, rien ne nous prouve qu'il se lie directement à l'ADN, et rien ne nous le prouvera tant qu'il n'y aura pas de validation *in vivo* (via les essais luciférase). C'est d'ailleurs pour cela que *STOX1* n'est toujours pas référencé comme facteur de transcription dans les bases de données. Certains éléments suggèreraient même une action de *STOX1* sur l'état chromatinien (article en cours de publication), ce qui pourrait expliquer les nombreuses modifications transcriptomiques observées lors de sa surexpression dans des cellules trophoblastiques. Mais à ce moment-là, pourquoi STRE2 ne fonctionne-t-il pas en retard sur gel car normalement, elle n'a pas été sélectionnée pour rien ? Est-ce que la protéine reconnaissant STRE2 ne s'y fixe qu'en présence d'un partenaire attaché à STRE1 ?

Pour poursuivre ce travail, la première approche serait de tester les promoteurs de gènes très dérégulés par *STOX1* (issus des expériences de transcriptomique) en utilisant MEME. Nous pourrons ainsi comparer ces séquences avec celles que nous avons trouvées et analyser leur environnement. Pour aller plus loin dans nos résultats de retard sur gel, il serait bon de faire l'expérience en utilisant de la protéine purifiée pour être sûr que l'interaction est indirecte. Il peut être également intéressant de tester le retard sur gel avec des sondes plus longues contenant les deux séquences consensus trouvées STRE1 et STRE2. Si ces premières expériences nous orientent vers la recherche de partenaires, nous pouvons envisager des expériences de co-immunoprécipitation suivies de spectrométrie de masse afin d'essayer de les déterminer. Concernant les essais luciférase, je n'ai eu le temps de faire qu'une expérience ce qui est trop peu pour conclure. Même si les résultats sont encourageants, il faudra en parallèle continuer ces expériences.

Ce travail sur *STOX1* s'intégrait donc au sein d'études menées pour comprendre de façon intégrée la fonction de ce gène majeur de la placentation, et sa dérégulation dans les cas de prééclampsie.

CONCLUSION

La diversité des réponses génétiques à un environnement, que l'on peut observer dans les cellules des mammifères, est en grande partie le résultat de facteurs de transcription qui régissent la façon dont les gènes sont transcrits et dont l'ARN polymérase II est recrutée. Grâce à leurs mécanismes complexes, les facteurs de transcription contrôlent des aspects importants du développement et du métabolisme des mammifères. Le répertoire des domaines de liaison à l'ADN, les interactions avec d'autres facteurs de transcription et des protéines associées à la chromatine, et les moyens de modifier la transcription s'agrandissent d'année en année, complexifiant ainsi le modèle initial des scientifiques de la lecture de l'information génétique par les cellules eucaryotes.

La quantité de données toujours croissante présente de nouveaux défis et ouvre de nouveaux horizons dans l'analyse des facteurs de transcription, de leurs mécanismes complexes et variés, et de leurs fonctions, et ce pas seulement chez les mammifères, mais aussi chez les plantes, chez les champignons et chez les procaryotes. En outre, dans cette thèse, je n'ai pas considéré les complexités supplémentaires de remodelage de la chromatine, de la régulation de la stabilité de l'ARNm, et du contrôle de la traduction, qui participent à la variété des réponses génétiques aux signaux environnementaux que l'on peut observer chez les mammifères.

Les études poussées de ces protéines et la compréhension de leur mécanisme d'action permettent d'entrevoir aujourd'hui les premières applications scientifiques : l'induction de cellules souches pluripotentes par l'équipe de Yamanaka a été une découverte révolutionnaire qui a complètement changé notre vision du développement et de la spécialisation cellulaire et qui ouvrent de nouvelles voies de thérapeutiques. Les résultats prometteurs sur les nouveaux champs d'applications de ces protéines aux actions diverses nous font persévérer dans la recherche et l'étude plus poussée de leurs mécanismes d'action, pour faire encore et toujours avancer la science au service de l'humanité.

BIBLIOGRAPHIE

Alberts B, Johnson A, Lewis J (2011). Biologie moléculaire de la cellule, cinquième. ed. Médecine Sciences Publications, Paris.

Andersson LS, Larhammar M, Memic F, Wootz H, Schwochow D, Rubin C-J, *et al.* (2012). Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. *Nature*, **488**, 642–646.

Bailey TL, Elkan C (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, **2**, 28–36.

Benayoun BA, Caburet S, Dipietromaria A, Bailly-Bechet M, Batista F, Fellous M, *et al.* (2008). The identification and characterization of a FOXL2 response element provides insights into the pathogenesis of mutant alleles. *Hum. Mol. Genet.*, **17**, 3118–3127.

Buckingham KJ, McMillin MJ, Brassil MM, Shively KM, Magnaye KM, Cortes A, *et al.* (2013). Multiple mutant T alleles cause haploinsufficiency of Brachyury and short tails in Manx cats. *Mamm. Genome*, **24**.

Cnattingius S, Reilly M, Pawitan Y, Lichtenstein P (2004). Maternal and fetal genetic factors account for most of familial aggregation of preeclampsia: a population-based Swedish cohort study. *Am. J. Med. Genet. A*, **130A**, 365–371.

Collas P, Dahl JA (2008). Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Front. Biosci.*, **13**, 929–943.

Crick F (1970). Central dogma of molecular biology. *Nature*, **227**, 561–563.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, *et al.* (2012). Landscape of transcription in human cells. *Nature*, **489**, 101–108.

Doridot L, Passet B, Méhats C, Rigourd V, Barbaux S, Ducat A, *et al.* (2013). Preeclampsia-Like Symptoms Induced in Mice by Fetoplacental Expression of STOX1 Are Reversed by Aspirin Treatment. *Hypertension*, **61**, 662–668.

Drögemüller C, Karlsson EK, Hytönen MK, Perloski M, Dolf G, Sainio K, *et al.* (2008). A mutation in hairless dogs implicates FOXI3 in ectodermal development. *Science*, **321**, 1462.

Dunker AK, Uversky VN (2010). Drugs for “protein clouds”: targeting intrinsically disordered transcription factors. *Curr Opin Pharmacol*, **10**, 782–788.

ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, *et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Founds SA, Conley YP, Lyons-Weiler JF, Jeyabalan A, Hogge WA, Conrad KP (2009). Altered global gene expression in first trimester placentas of women destined to develop preeclampsia. *Placenta*, **30**, 15–24.

Franklin RE, Gosling RG (1953). Molecular configuration in sodium thymonucleate. *Nature*, **171**, 740–741.

Fuchs R, Ellinger I (2004). Endocytic and transcytotic processes in villous syncytiotrophoblast: role in nutrient transport to the human fetus. *Traffic*, **5**, 725–738.

Gommans WM, Haisma HJ, Rots MG (2005). Engineering zinc finger protein transcription factors: the therapeutic relevance of switching endogenous gene expression on or off at command. *J. Mol. Biol.*, **354**, 507–519.

Griffiths AJF, Wessler SR, Caroll SB (2013). Introduction à l’analyse génétique, 6ème ed. de boeck, Bruxelles.

Gurdon JB (1962). The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *J Embryol Exp Morphol*, **10**, 622–640.

Haworth K, Putt W, Cattanach B, Breen M, Binns M, Lingaas F, *et al.* (2001). Canine homolog of the T-box transcription factor T; failure of the protein to bind to its DNA target leads to a short-tail phenotype. *Mamm. Genome*, **12**, 212–218.

Hytönen MK, Grall A, Hédan B, Dréano S, Seguin SJ, Delattre D, *et al.* (2009). Ancestral T-box mutation is present in many, but not all, short-tailed dog breeds. *J. Hered.*, **100**, 236–240.

Kaplan J-C, Delpech M (2007). Biologie moléculaire et médecine, troisième. ed, *De la biologie à la clinique*,. Médecine Sciences Flammarion, Paris.

Kaufmann P, Black S, Huppertz B (2003). Endovascular trophoblast invasion: implications for the pathogenesis of intrauterine growth retardation and preeclampsia. *Biol. Reprod.*, **69**, 1–7.

Kijas JW, Ortiz JS, McCulloch R, James A, Brice B, Swain B, *et al.* (2013). Genetic diversity and investigation of polledness in divergent goat populations using 52 088 SNPs. *Anim. Genet.*, **44**, 325–335.

Levine M, Tjian R (2003). Transcription regulation and animal diversity. *Nature*, **424**, 147–151.

Luscombe NM, Austin SE, Berman HM, Thornton JM (2000). An overview of the structures of protein-DNA complexes. *Genome biology*, **1**, reviews001.

Maniatis T, Goodbourn S, Fischer JA (1987). Regulation of inducible and tissue-specific gene expression. *Science*, **236**, 1237–1245.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, *et al.* (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–110.

Mellersh CS, McLaughlin B, Ahonen S, Pettitt L, Lohi H, Barnett KC (2009). Mutation in HSF4 is associated with hereditary cataract in the Australian Shepherd. *Vet Ophthalmol*, **12**, 372–378.

Mellersh CS, Pettitt L, Forman OP, Vaudin M, Barnett KC (2006). Identification of mutations in HSF4 in dogs of three different breeds with hereditary cataracts. *Vet Ophthalmol*, **9**, 369–378.

Metallo SJ (2010). Intrinsically disordered proteins are potential drug targets. *Curr Opin Chem Biol*, **14**, 481–488.

Miller J, McLachlan AD, Klug A (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J.*, **4**, 1609–1614.

Pailhoux E, Vigier B, Chaffaux S, Servel N, Taourit S, Furet JP, *et al.* (2001). A 11.7-kb deletion triggers intersexuality and polledness in goats. *Nat. Genet.*, **29**, 453–458.

Pailhoux E, Vigier B, Schibler L, Cribiu EP, Cotinot C, Vaiman D (2005). Positional cloning of the PIS mutation in goats and its impact on understanding mammalian sex-differentiation. *Genet. Sel. Evol.*, **37 Suppl 1**, S55–64.

Petersen JL, Mickelson JR, Rendahl AK, Valberg SJ, Andersson LS, Axelsson J, *et al.* (2013). Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS Genet.*, **9**, e1003211.

Phillips T, Hoopes L (2008). Transcription Factors and Transcriptional Control. *Nature Education 1(1)*,.

Pray LA (2008). Discovery of DNA structure and function: Watson and Crick. *Nature Education 1(1)*,.

Qian Z, Cai Y-D, Li Y (2006). Automatic transcription factor classifier based on functional domain composition. *Biochemical and Biophysical Research Communications*, **347**, 141–144.

R. Hughes T (Ed.) (2011). A Handbook of Transcription Factors, *Springer*,. Timithy R. Hughes, Dordrecht Heidelberg London New-York.

Rigourd V, Chauvet C, Chelbi ST, Rebourcet R, Mondon F, Letourneur F, *et al.* (2008). STOX1 overexpression in choriocarcinoma cells mimics transcriptional alterations observed in preeclamptic placentas. *PLoS ONE*, **3**, e3905.

Rigourd V, Chelbi S, Chauvet C, Rebourcet R, Barbaux S, Bessières B, *et al.* (2009). Re-evaluation of the role of STOX1 transcription factor in placental development and preeclampsia. *J. Reprod. Immunol.*, **82**, 174–181.

Rosier F (2012). Vers la fin de l'"ADN poubelle" *In: Le Monde.fr* [en ligne]. [http://www.lemonde.fr/sciences/article/2012/09/06/vers-la-fin-de-l-adn-poubelle_1756718_1650684.html] (Consultation le 14/10/13).

Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, *et al.* (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, **131**, 861–872.

Takahashi K, Yamanaka S (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**, 663–676.

The 2012 Nobel Prize in Physiology or Medicine - Press Release (2012). *In: Nobelprize.org* [en ligne]. [http://www.nobelprize.org/nobel_prizes/medicine/laureates/2012/press.html] (Consultation le 8/9/13).

Tjian R (1995). Molecular machines that control genes. *Sci. Am.*, **272**, 54–61.

Tsai KL, Noorai RE, Starr-Moss AN, Quignon P, Rinz CJ, Ostrander EA, *et al.* (2012). Genome-wide association studies for multiple diseases of the German Shepherd Dog. *Mamm. Genome*, **23**, 203–211.

Van Dijk M, Mulders J, Poutsma A, Könst AAM, Lachmeijer AMA, Dekker GA, *et al.* (2005). Maternal segregation of the Dutch preeclampsia locus at 10q22 with a new member of the winged helix gene family. *Nat. Genet.*, **37**, 514–519.

Voorbij AMWY, van Steenbeek FG, Vos-Loohuis M, Martens EECP, Hanson-Nilsson JM, van Oost BA, *et al.* (2011). A contracted DNA repeat in LHX3 intron 5 is associated with aberrant splicing and pituitary dwarfism in German shepherd dogs. *PLoS ONE*, **6**, e27940.

Watson JD, Crick FHC (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.

Wiener DJ, Gurtner C, Panakova L, Mausberg T-B, Müller EJ, Drögemüller C, *et al.* (2013). Clinical and histological characterization of hair coat and glandular tissue of Chinese crested dogs. *Vet. Dermatol.*, **24**, 274–e62.

Wilkins MHF, Stokes AR, Wilson HR (1953). Molecular structure of deoxypentose nucleic acids. *Nature*, **171**, 738–740.

Yan C, Higgins PJ (2013). Drugging the undruggable: transcription therapy for cancer. *Biochim. Biophys. Acta*, **1835**, 76–85.

Zaehres H, Schöler HR (2007). Induction of pluripotency: from mouse to human. *Cell*, **131**, 834–835.

Zhang H-M, Chen H, Liu W, Liu H, Gong J, Wang H, *et al.* (2012). AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res.*, **40**, D144–149.

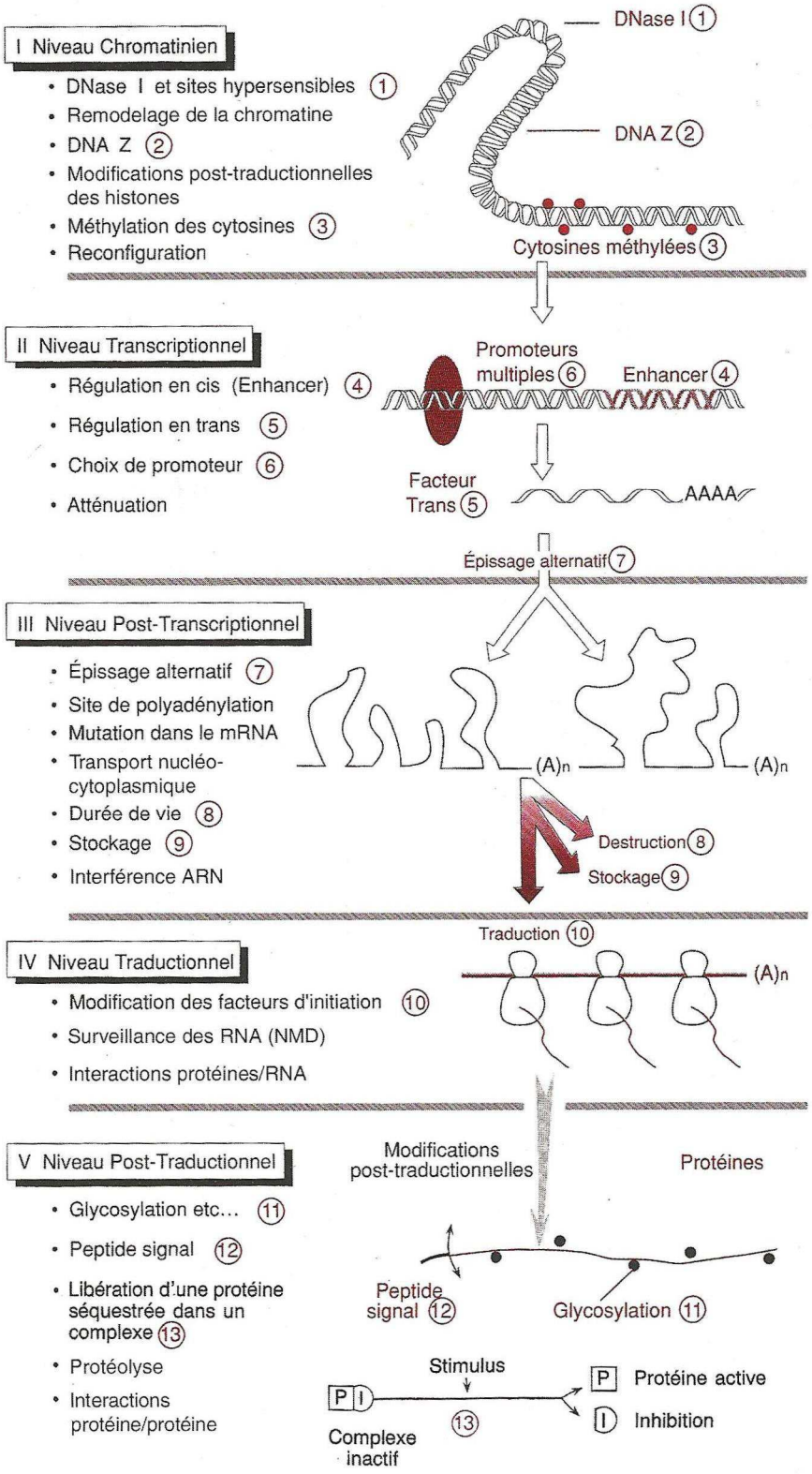
Zhou P, Sun LJ, Dötsch V, Wagner G, Verdine GL (1998). Solution structure of the core NFATC1/DNA complex. *Cell*, **92**, 687–696.

LISTE DES SITES WEB

- *AnimalTFDB (base de données de facteurs de transcription)* [http://www.bioguo.org/AnimalTFDB/family_index.php] (Consulté le 27/08/13)
- *Club du Chihuahua, du Coton de tulear et des Exotiques (CCCE)* [http://www.ccce.org/races.html] (Consultation le 25/08/13)
- *College National des Gynécologues et Obstétriciens Français (CNGOF)* [http://www.cngof.asso.fr/index.html] (Consultation le 10/10/13)
- *ENCODE (Encyclopedia Of DNA Elements), page du projet* [http://www.encodeproject.org/ENCODE/] (Consultation le 20/10/13)
- *Ensembl* [http://www.ensembl.org/index.html] (Consultation le 05/09/13)
- *Eukarytic promotor Database* [http://epd.vital-it.ch/] (Consultation le 05/09/13)
- *Eurogentec* [http://www.eurogentec.com/eu-home.html] (Consultation le 12/10/13)
- *InfoVeto* [http://www.infoveto.com/race/welsh-corgi/] (Consultation le 25/08/13)
- *Life Technologies™* [http://www.lifetechnologies.com/] (Consultation le 12/10/13)
- *MEME* (*Multiple Em for Motif Elicitation*) [http://meme.nbcr.net/] (Consultation le 15/05/13)
- *MGI (Mouse Genome Informatics)* [http://www.informatics.jax.org/] (Consultation le 23/08/13)
- *NCBI (National Centre for Biotechnology Information)* [http://www.ncbi.nlm.nih.gov.gate2.inist.fr/] (Consultation le 05/09/13)
- *OMIA* (*Online Mendelian Inheritance in Animals*) [http://omia.angis.org.au/home/] (Consultation le 23/08/13)
- *OMIM®* (*Online Mendelian Inheritance in Man*) [http://www.omim.org/] (Consultation le 23/08/13)
- *PDB (Protein Data Bank)* [http://www.pdb.org/pdb/home/home.do] (Consultation le 05/09/13)
- *Pfam 27.0 (Protein families, mars 2013, base de données comprenant 14 831 familles protéiques, permettant de faire des alignements de séquences)* [http://pfam.sanger.ac.uk/] (Consultation le 05/09/13)
- *Prix Nobel (site officiel)* [http://www.nobelprize.org/] (Consultation le 27/08/13)
- *Promega* [http://france.promega.com/] (Consultation le 12/10/13)
- *PSIPRED (Protein Structure Prediction server)* [http://bioinf.cs.ucl.ac.uk/psipred/] (Consultation le 05/09/13)
- *SIFT (Sorting Intolerant From Tolerant)* [http://blocks.fhcrc.org/sift/sift.html] (Consultation le 05/09/13)
- *TRANSFAC (base de données de facteurs de transcription)* [http://www.biobase-international.com/gene-regulation] (Consultation le 27/08/13)
- *Thermo Scientific* [http://www.thermoscientific.com/] (Consultation le 12/10/13)

ANNEXES

Annexe 1 : Régulation de l'expression génique chez les eucaryotes



D'après (Kaplan et Delpech, 2007). Chez les eucaryotes, il existe une multiplicité de niveaux successifs de régulation de la transcription, qui vont du niveau chromatinien (niveau élevé de régulation) au niveau post-traductionnel (dernier stade possible de régulation).

Annexe 2 : Tableau de maladies héréditaires dues à un facteur de transcription chez l'homme et la souris

Gène *Anciennes appellations* **Chromosome chez l'homme**/chromosome chez la souris	Maladie associée chez l'homme
Aire (*Autoimmune regulator*) *autoimmune polyendocrinopathy candidiasis ectodermal dystrophy* **21**/10	Polyendocrinopathie auto-immune de type I
Ar (*Androgen receptor*) **X**/X	Syndrome d'insensibilité aux androgènes + Amyotrophie bulbo-spinale liée à l'X ou maladie de Kennedy
Arx (*Aristaless related homeobox*) **X**/X	L'encéphalopathie épileptique infantile précoce ou syndrome d'Ohtahara + Lissencéphalie liée à l'X
Cebpa (*CCAAT/enhancer binding protein* (C/EBP) *alpha*) *Cebp, C/ebpalpha, C/EBP alpha* **19**/7	Leucémie Aiguë Myéloïde
Crx (*cone-rod homeobox containing gene*) *Crx1* **19**/7	Amaurose congénitale de Leber
Egr2 (*early growth response 2*) *Egr-2, Krox20, Krox-20, NGF1-B, Zfp-25* **10**/10	Neuropathie hypomyélinisante congénitale
Foxc2 (*forkhead box C2*) *Fkh14, Hfhbf3, Mfh1, MFH-1* **16**/8	Syndrome lymphœdème-distichiasis
Foxe1 (*forkhead box E1*) *thyroid transcription factor 2, Titf2* **9**/4	Hypothyroïdie congénitale avec cheveux bouclés et fente palatine
Foxe3 (*forkhead box E3*) *FREAC8, rct* **1**/4	Dysgénésie mésenchymateuse du segment antérieur de l'œil
Foxi1 (*forkhead box I1*) *Fkh10, Hfh3, HFH-3* **5**/11	Syndrome de Pendred
Foxl2 (*forkhead box L2*) *Pfrk* **3**/9	Syndrome blépharophimosis-ptosis-épicanthus inversus
Foxn1 (*forkhead box N1*) *D11Bhm185e, Hfh11, whn* **17**/11	Immunodéficience sévère en cellules T - alopécie congénitale - dystrophie des ongles
Foxp2 (*forkhead box P2*) *2810043D05Rik, D0Kist7* **7**/6	Trouble du langage et de la parole de type 1
Foxp3 (*forkhead box P3*) *JM2, scurfin* **X**/X	Syndrome IPEX (Immuno-dérégulation, Poly-endocrinopathie, Entéropathie, liée a l'X)
Gata3 (*GATA binding protein 3*) *Gata-3, jal* **10**/2	Syndrome de Bartter (syndrome « reins-oreilles » avec hypoparathyroidisme, surdité sensorineurale, et insuffisance rénale)
Gata4 (*GATA binding protein 4*) *Gata-4* **8**/14	Communication inter-auriculaire de type 2
Gcm2 (*glial cells missing homolog 2 (Drosophila)*) *Gcm1-rs2* **6**/13	Hypoparathyroïdie isolée familiale

Gène *Anciennes appellations* **Chromosome chez l'homme**/chromosome chez la souris	Maladie associée chez l'homme
Gfi1 (*growth factor independent 1*) *Gfi-1, Pal1, Pal-1* **1**/5	Neutropénie congénitale sévère de type 2, autosomique dominante
Gli3 (GLI-Kruppel family member GLI3) *Bph, brachyphalangy* **7**/13	Syndrome de Greig + Syndrome de Pallister-Hall
Glis2 (*GLIS family zinc finger 2*) *MGC:36775, Nkl* **16**/16	Néphronophtise de type 7
Glis3 (*GLIS family zinc finger 3*) *4833409N03Rik, E330013K21Rik* **9**/19	Diabète sucré néonatal avec hypothyroïdie congénitale
Hmx1 (*H6 homeobox 1*) *Nkx5-3* **4**/5	Syndrome oculo-auriculaire
Hnf1a (HNF1 homeobox A) *hepatocyte nuclear factor 1, Hnf-1, HNF1, HNF1[a], Hnf1alpha, HNF1-alpha, LFB1, Tcf1* **12**/5	Diabète sucré non-insulino-dépendant + Diabète du jeune de type 3
Hnf4a (*hepatic nuclear factor 4, alpha*) *Hnf4, HNF-4, HNF4 alpha, MODY1, Nr2a1, Nuclear receptor 2A1, Tcf14, Tcf4* **20**/2	Diabète du jeune de type 1
Hesx1 (*homeobox gene expressed in ES cells*) *HES-1, Rpx* **3**/14	Dysplasie septo-optique
Hoxa1 (*homeobox A1*) *early retinoic acid, ERA1, Hox-1.6* **7**/6	Syndrome de dysgénésie de tronc cérébral d'Athabascan
Hoxa13 (*homeobox A13*) *Hox-1.10* **7**/6	Syndrome main-pied-utérus
*Hoxd1*3 (*homeobox D13*) *Hox-4.8, spdh* **2**/2	Sympolydactylie de type 1
Hr (*Hairless*) *ba, bldy, N, rh, rh-bmh* **8**/14	Pelade universelle congénitale + Atrichie avec lésions papulaires + Hypotrichose de type 4
Hsf4 (heat shock transcription factor 4) *ldis1* **16**/8	Cataracte de Marner
Irf6 (*interferon regulatory factor 6*) *E230028I05Rik* **1**/1	Le syndrome du pterygium poplité + Syndrome Van Der Woude de type 1
Lhx3 (*LIM homeobox protein 3*) *Lim3, mLim-3, P-LIM* **9**/2	Déficit hypophysaire combiné multiple de type 3
Lmx1b (*LIM homeobox transcription factor 1 bêta*) *LMX1.2* **9**/2	Syndrome ongles-rotule
Maf (*avian musculoaponeurotic fibrosarcoma (v-maf) AS42 oncogene homolog*) *2810401A20Rik, A230108G15Rik, c-maf* **16**/8	Cataracte de type 21

Gène *Anciennes appellations* **Chromosome chez l'homme**/chromosome chez la souris	Maladie associée chez l'homme
Mecp2 (*methyl CpG binding protein 2*) *1500041B07Rik, D630021H01Rik, Mbd5, WBP10* **X**/X	Syndrome de Rett
Men1 (*multiple endocrine neoplasia 1*) *menin* **11**/19	Néoplasie endocrinienne multiple de type I
Mitf (*microphthalmia-associated transcription factor*) *BCC2, bHLHe32, Gsfbcc2, mi, wh* **3**/6	Albinisme oculaire avec surdité sensorielle tardive + Syndrome de Tietz + Syndrome de Waardenburg de type 2A
Msx1 (*msh homeobox 1*) *Hox7, Hox-7, Hox7.1, msh, muscle-segment homeobox* **4**/5	Fente oro-faciale de type 5 + Agénésie dentaire sélective de type 1
Msx2 (*msh homeobox 2*) *Hox8, Hox-8, Hox8.1* **5**/13	Craniosynostose de type 2 + Foramen pariétal
Myc (myelocytomatosis oncogene) *bHLHe39, c-myc, Myc2, Niard, Nird* **8**/15	Lymphome de Burkitt
Nr0b1 (*nuclear receptor subfamily 0, group B, member 1*) *Ahc, Ahch, AHX, Dax1, DAX-1* **X**/X	Hypoplasie congénitale des surrénales
Nr2e3 (*nuclear receptor subfamily 2, group E, member 3*) *photoreceptor-specific nuclear receptor, Pnr, RNR* **15**/9	Syndrome d'augmentation des cônes bleus ou syndrome de Goldman-Favre
Nr3c2 (*nuclear receptor subfamily 3, group C, member 2*) *aldosterone receptor, mineralocorticoid receptor, Mlr, MR* **4**/8	Pseudohypoaldostéronisme de type I, autosomique dominant
Pax2 (*paired box gene 2*) *Opdc, Pax-2* **10**/19	Syndrome papillo-rénal
Pax3 (*paired box gene 3*) *Pax-3* **2**/1	Rhabdomyosarcome de type 2 + Syndrome de Waardenburg de type 1
Pax6 (*paired box gene 6*) *1500038E17Rik, AEY11, Dey, Dickie's small eye, Gsfaey11, Pax-6* **11**/2	Aniridie de type 2 + Kératite héréditaire + Glaucome congénital de Peters ou anomalie de Peters
Pax8 (*paired box gene 8*) **2**/2	Hypothyrodië congénitale sans goître de type 2
Pax9 (*paired box gene 9*) *Pax-9* **14**/12	Agénésie dentaire sélective de type 3
Pdx1 (pancreatic and duodenal homeobox 1) *IDX-1, Ipf1, IPF-1, Mody4, pdx-1, STF-1* **13**/5	Diabète sucré du jeune de type 4
Phox2b (*paired-like homeobox 2b*) *Dilp1, GENA 269, NBPhox, Phox2b, Pmx2b* **4**/5	Syndrome d'Ondine ou syndrome d'hypoventilation alvéolaire centrale congénitale ou ondinisme
Pitx1 (*paired-like homeodomain transcription factor 1*) *Bft, Potx, P-OTX, Ptx1* **5**/13	Pieds bots congénitaux ou déformations congénitales des pieds avec ou sans déficience des longs os et/ou polydactylie miroir
Pitx2 (*paired-like homeodomain transcription factor 2*) *Brx1, Brx1a, Brx1b, Munc30, Otlx2, Pitx2a, Pitx2b, Pitx2c, Ptx2, Rieg, solurshin* **4**/3	Syndrome d'Axenfeld-Rieger de type 1

Gène *Anciennes appellations* **Chromosome chez l'homme**/chromosome chez la souris	Maladie associée chez l'homme
Pitx3 (paired-like homeodomain transcription factor 3) *Ptx3* **10**/19	Dysgénésie mésenchymateuse du segment antérieur de l'œil
Pou3f4 (*POU domain, class 3, transcription factor 4*) *Brn4, BRN-4, Otf9* **X**/X	Surdité liée à l'X de type 2
Pparg (*peroxisome proliferator activated receptor gamma*) *Nr1c3, PPARgamma, PPAR-gamma, Ppar-gamma2, PPARgamma2* **3**/6	Lipodystrophie partielle familiale de type 3
Prop1 (*paired like homeodomain factor 1*) *Prop-1, prophet of Pit1, prophet of Pit-1* **5**/11	Déficit hypophysaire combiné multiple de type 2
Runx2 (*runt related transcription factor 2*) *AML3, Cbfa1, Osf2, PEBP2aA, PEBP2 alpha A, Pebpa2a, polyomavirus enhancer binding factor 2 (PEBP2), SL3-3 enhancer factor 1* **6**/17	Dysplasie cléidocrânienne
Sall1 (*sal-like 1 (Drosophila)*) *Msal-3* **16**/8	Syndrome de Townes-Brocks
Sall4 (sal-like 4 (Drosophila)) *5730441M18Rik, C330011P20Rik, Tex20* **20**/2	Anomalie de Duane-Radial Ray ou syndrome de Okihiro
Sim1 (*single-minded homolog 1 (Drosophila)*) *bHLHe14* **6**/10	Obésité sévère
Six3 (sine oculis-related homeobox 3) *E130112M24Rik* **2**/17	Holoprosencéphalie de type 2
Smad4 (*SMAD family member 4*) *D18Wsu70e, Dpc4, DPC4, Madh4, Smad 4* **18**/18	Polypose gastro-intestinale juvénile
Smad9 (SMAD family member 9) *MADH6, Madh9, SMAD8A, SMAD8B* **13**/3	Hypertension artérielle pulmonaire primitive de type 1
Snai2 (*snail homolog 2 (Drosophila)*) *Slug, Slugh, Snail2* **8**/16	Syndrome de Waardenburg de type 2D
Sox9 (*SRY-box containing gene 9*) *2010306G03Rik* **17**/11	Dysplasie campomélique
Sox18 (*SRY-box containing gene 18*) *Ragl, Sry-related HMG-box gene 18* **20**/2	Syndrome hypotrichose - lymphoedème - télangiectasie
Tbx1 (*T-box 1*) **22**/16	Syndrome de DiGeorge ou syndrome vélocardiofacial
Tbx19 (*T-box 19*) *D1Ertd754e, Tpit* **1**/1	Déficit congénital isolé en ACTH
Tbx22 (*T-box 22*) *D230020M15Rik* **X**/X	Fente palatine et ankyloglossie liées à l'X
Tbx3 (*T-box 3*) *D5Ertd189e* **12**/5	Syndrome ulnaire-mammaire
Tbx5 (*T-box 5*) **12**/5	Syndrome de Holt-Oram

Gène *Anciennes appellations* **Chromosome chez l'homme**/chromosome chez la souris	Maladie associée chez l'homme
Tfap2b (*transcription factor AP-2 bêta*) *AP-2(beta), E130018K07Rik, Tcfap2b* **6**/1	Syndrome de Char
Thrb (*thyroid hormone receptor bêta*) *c-erbAbeta, Nr1a2, T3R[b], T3Rbeta, Thrb1, Thrb2, TR beta* **3**/14	Résistance généralisée aux hormones thyroïdiennes, autosomique récessif
Trp53 (*transformation related protein 53*) *p44, p53* **17**/11	Cancer du sein + Syndrome de Li-Fraumeni de type 1 + Cancer du pancréas
Trp63 (*transformation related protein 63*) *deltaNp63, KET protein, p51/p63, p63, p73L, TAp63, Trp53rp1* **3**/16	Ankyloblépharon - anomalies ectodermiques - fente labiopalatine ou syndrome de Hay-Wells + Syndrome Ectrodactylie - dysplasie ectodermique - fente labiopalatine de type 3
Trps1 (*trichorhinophalangeal syndrome I (human)*) *D15Ertd586e* **8**/15	Syndrome tricho-rhino-phalangien de type 1
Twist1 (*twist basic helix-loop-helix transcription factor 1*) *bHLHa38, charlie chaplin, M-Twist, pdt, Pluridigite, Ska10, Ska* **7**/12	Syndrome de Saethre-Chotzen
Vdr (*vitamin D receptor*) *Nr1i1* **12**/15	Osteoporose + Rachitisme hypocalcémique vitamine D-dépendant de type 2A
Zeb1 (*zinc finger E-box binding homeobox 1*) *3110032K11Rik, AREB6, [delta]EF1, MEB1, Nil2, Tcf18, Tcf8, Tw, ZEB, Zfhep, Zfhx1a, Zfx1a* **10**/18	Dystrophie postérieure polymorphe ou dystrophie de Schlichting
Zeb2 (*zinc finger E-box binding homeobox 2*) *9130203F04Rik, D130016B08Rik, mKIAA0569, SIP1, Zfhx1b, Zfx1b* **2**/2	Syndrome de Mowat-Wilson
Zfpm2 (zinc finger protein, multitype 2) *B330005D23Rik, FOG2, FOG-2* **8**/15	Tétralogie de Fallot
Zic2 (*zinc finger protein of the cerebellum 2*) *GENA 29, Ku, odd-paired homolog* **13**/14	Holoprosencéphalie de type 5
Zic3 (*zinc finger protein of the cerebellum 3*) **X**/X	Hétérotaxie viscérale liée à l'X de type 1
Whsc1 (*Wolf-Hirschhorn syndrome candidate 1 (human)*) *5830445G22Rik, 9430010A17Rik, C130020C13Rik, D030027O06Rik, D930023B08Rik, mKIAA1090, Whsc1l* **4**/5	Syndrome de Wolf-Hirschhorn
Wt1 (*Wilms tumor 1 homolog*) *D630046I19Rik, Wt-1* **11**/2	Syndrome de Denys-Drash

Annexe 3 : Transcrits et séquences protéiques des différentes isoformes de *STOX1*

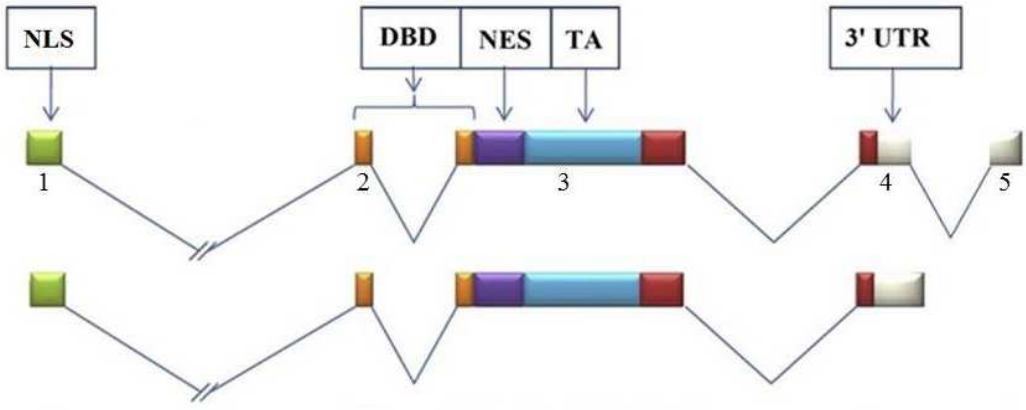
D'après les bases de données NCBI (*National Centre for Biotechnology Information* : http://www.ncbi.nlm.nih.gov.gate2.inist.fr/) et *Ensembl* (http://www.ensembl.org/index.html), par épissage alternatif de l'exon 3 (complet ou partiel) et de l'exon 5, cinq transcrits différents sont générés et codent pour quatre isoformes (A, B, C, D). Les couleurs identiques à l'intérieur des exons indiquent des séquences de protéines identiques.

Légende :

NLS (*Nuclear Localization Sequence*) : Séquence de localisation nucléaire
DBD (*DNA Binding Domain*) : Site de fixation à l'ADN
NES (*Nuclear Export Signal*) : Signal d'export nucléaire.
TA (*Transactivator domain*) : domaine transactivateur.

Isoforme A

Transcrits : deux transcrits légèrement différents codent pour la même isoforme



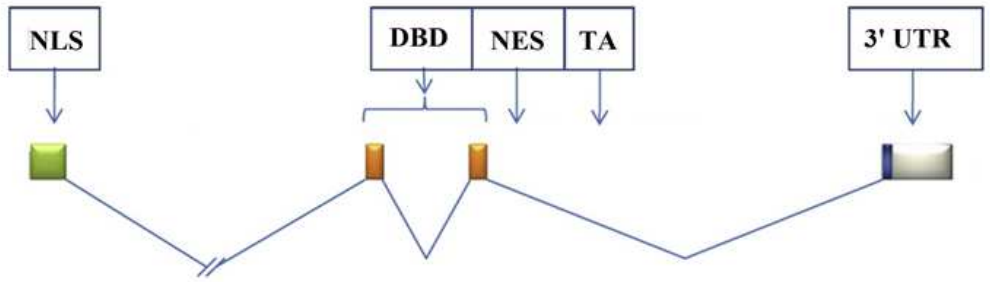
Séquence protéique :

MARPVQLAPGSLALVLCRLEAQKAAGAAEEPGGRAVFRAFRRANARCFWNARLARAASR
LAFQGWLRRGVLLVRAPPACLQVLRDAWRRRALRPPRGFRIRAVGDVFPVQMNPITQSQF
VPLGEVLCCAISDMNTAQIVVTQESLLERLMKHYPGIAIPSEDILYTTLGTLIKERKIYHTGE
GYFIVTPQTYFITNTTTQENKRMLPSDESRLMPASMTYLVSMESCAESAQENAAPISHCQS
CQCFRDMHTQDVQEAPVAAEVTRKSHRGLGESVSWVQNGAVSVSAEHHICESTKPLPYTR
DKEKGKKFGFSLLWRSLSRKEKPKTEHSSFSAQFPPEEWPVRDEDDLDNIPRDVEHEIIKRI
NPILTVDNLIKHTVLMQKYEEQKKYNSQGTSTDMLTIGHKYPSKEGVKKRQGLSAKPQGQ
GHSRRDRHKARNQGSEFQPGSIRLEKHPKLPATQPIPRIKSPNEMVGQKPLGEITTVLGSHLI
YKKRISNPFQGLSHRGSTISKGHKIQKTSDLKPSQTGPKEKPFQKPRSLDSSRIFDGKAKEPY
AEQPNDKMEAESIYINDPTVKPINDDFRGHLFSHPQQSMLQNDGKCCPFMESMLRYEVYG
GENEVIPEVLRKSHSHFDKLGETKQTPHSLPSRGASFSDRTPSACRLVDNTIHQFQNLGLLD
YPVGVNPLRQAARQDKDSEELLRKGFVQDAETTSLENEQLSNDDQALYQNEVEDDDGAC
SSLYLEEDDISENDDLRQMLPGHSQYSFTGGSQGNHLGKQKVIERSLTEYNSTMERVESQV
LKRNECYKPTGLHATPGESQEPNLSAESCGLNSGAQFGFNYEEEPSVAKCVQASAPADERI
FDYYSARKASFEAEVIQDTIGDTGKKPASWSQSPQNQEMRKHFPQKFQLFNTSHMPVLAQ
DVQYEHSHLEGTENHSMAGDSGIDSPRTQSLGSNNSVILDGLKRRQNFLQNVEGTKSSQPL
TSNSLLPLTPVINV

➔ Il s'agit de la plus longue isoforme de STOX1.

Isoforme B

Transcrit :



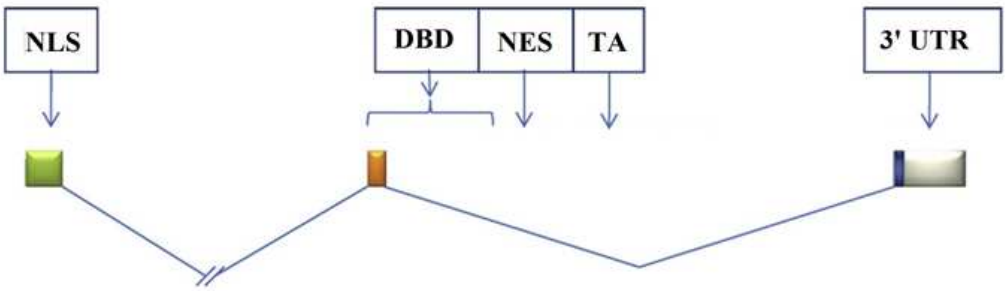
Séquence protéique :

MARPVQLAPGSLALVLCRLEAQKAAGAAEEPGGRAVFRAFRRANARCFWNARLARAASR
LAFQGWLRRGVLLVRAPPACLQVLRDAWRRRALRPPRGFRIRAVGDVFPVQMNPITQSQF
VPLGEVLCCAISDMNTAQIVVTQESLLERLMKHYPGIAIPSEDILYTTLGTLIKERKIYHTGE
GYFIVTPQTYFITNTTTQENKRMLPSDESRLMPASMTYLDTESGI

➔ Une modification post-transcriptionnelle de l'ARN utilisant un site donneur d'épissage au niveau de l'avant-dernier exon provoque un décalage du cadre de lecture d'où la formation d'une plus courte isoforme avec une extrémité C-terminale distincte de l'isoforme A.

Isoforme C

Transcrit :



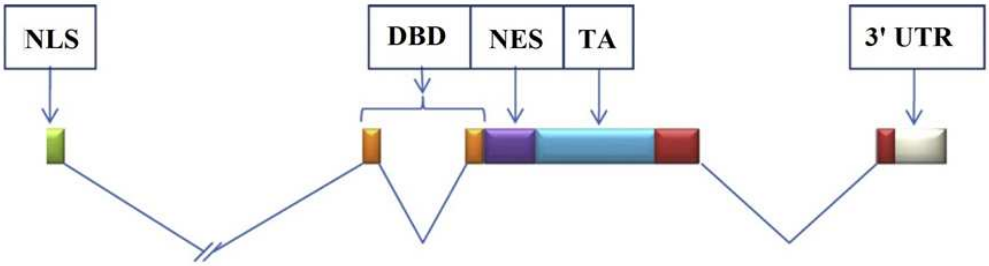
Séquence protéique :

MARPVQLAPGSLALVLCRLEAQKAAGAAEEPGGRAVFRAFRRANARCFWNARLARAASR
LAFQGWLRRGVLLVRAPPACLQVLRDAWRRRALRPPRGFRIRAVGDVFPVQMNPITQSQF
VPLGEVLCCAISDMNTAQIVVTQESLLERLMKHYPGHRVWDLIIQSFWMD

➔ Un épissage alternatif mène à la perte de l'extrémité 3' de l'ARN, d'où la formation d'une isoforme plus courte avec une extrémité C-terminale distincte de l'isoforme A.

Isoforme D

Transcrit :

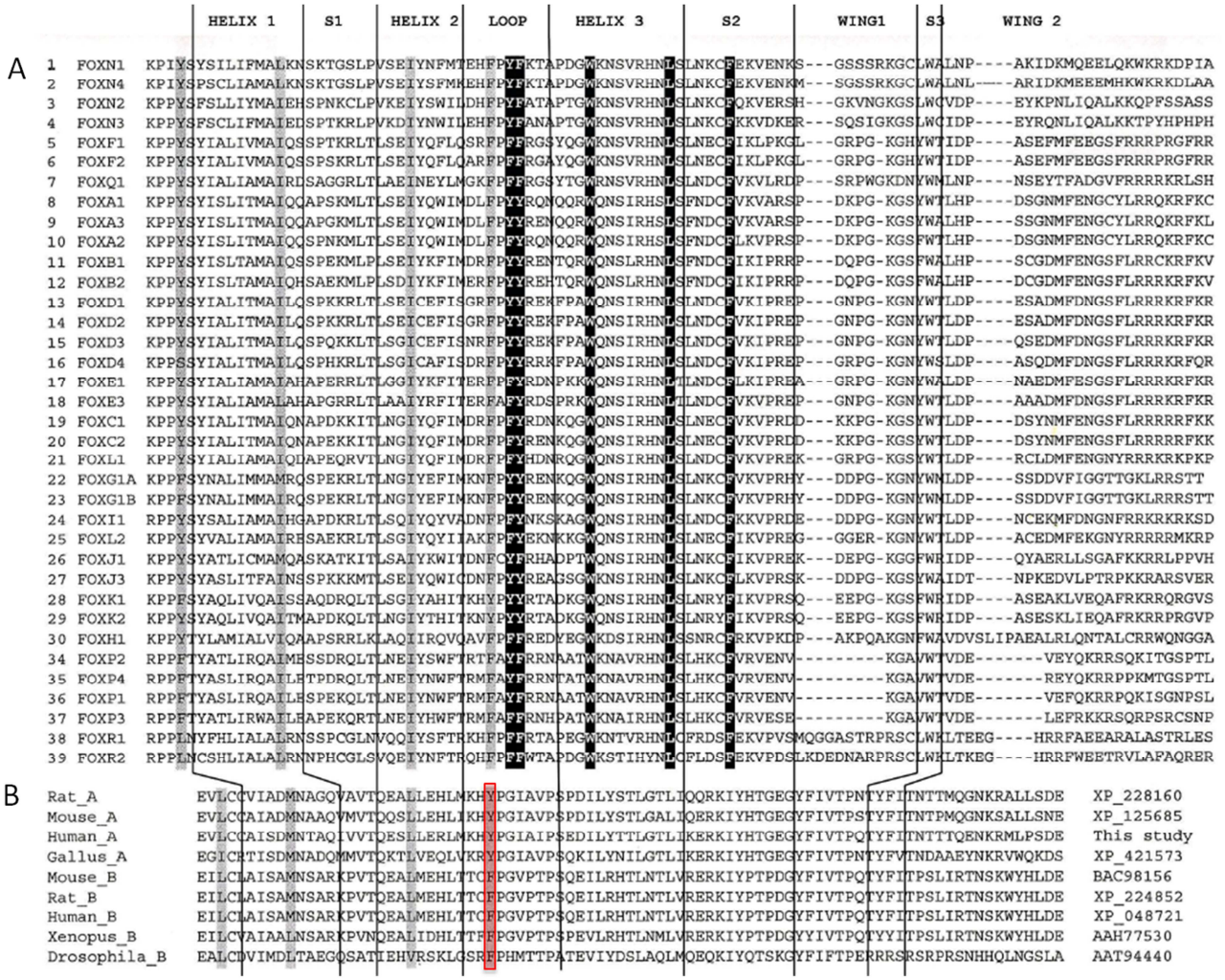


Séquence protéique :

MNPITQSQFVPLGEVLCCAISDMNTAQIVVTQESLLERLMKHYPGIAIPSEDILYTTLGTLIK
ERKIYHTGEGYFIVTPQTYFITNTTTQENKRMLPSDESRLMPASMTYLVSMESCAESAQEN
AAPISHCQSCQCFRDMHTQDVQEAPVAAEVTRKSHRGLGESVSWVQNGAVSVSAEHHICE
STKPLPYTRDKEKGKKFGFSLLWRSLSRKEKPKTEHSSFSAQFPPEEWPVRDEDDLDNIPRD
VEHEIIKRINPILTVDNLIKHTVLMQKYEEQKKYNSQGTSTDMLTIGHKYPSKEGVKKRQG
LSAKPQGQGHSRRDRHKARNQGSEFQPGSIRLEKHPKLPATQPIPRIKSPNEMVGQKPLGEI
TTVLGSHLIYKKRISNPFQGLSHRGSTISKGHKIQKTSDLKPSQTGPKEKPFQKPRSLDSSRIF
DGKAKEPYAEQPNDKMEAESIYINDPTVKPINDDFRGHLFSHPQQSMLQNDGKCCPFMES
MLRYEVYGGENEVIPEVLRKSHSHFDKLGETKQTPHSLPSRGASFSDRTPSACRLVDNTIHQ
FQNLGLLDYPVGVNPLRQAARQDKDSEELLRKGFVQDAETTSLENEQLSNDDQALYQNEV
EDDDGACSSLYLEEDDISENDDLRQMLPGHSQYSFTGGSQGNHLGKQKVIERSLTEYNSTM
ERVESQVLKRNECYKPTGLHATPGESQEPNLSAESCGLNSGAQFGFNYEEEPSVAKCVQAS
APADERIFDYYSARKASFEAEVIQDTIGDTGKKPASWSQSPQNQEMRKHFPQKFQLFNTSH
MPVLAQDVQYEHSHLEGTENHSMAGDSGIDSPRTQSLGSNNSVILDGLKRRQNFLQNVEG
TKSSQPLTSNSLLPLTPVINV

➔ L’extrémité 5’ de cet ARN est différente des autres, ce qui conduit à l’initiation de la traduction au niveau d’un AUG plus en aval mais dans le même cadre de lecture, d’où la formation d’une isoforme avec une extrémité N-terminale plus courte par rapport à l’isoforme A.

Annexe 4 : Alignement des séquences protéiques FOX avec la protéine STOX1



A. L'alignement du domaine de l'hélice ailée des protéines FOX chez l'homme montre une conservation des acides aminés hydrophiles (en gris) qui contrôlent la stabilité de l'aile 2 et ainsi contrôle la spécificité de liaison à l'ADN. En outre, toutes les protéines FOX suivent un schéma Y/FY/F-6-W-7-L-4-F absolument conservé (marqué en noir). Dans le cas des protéines FOXO, ce schéma est Y/FY/F-11-W-7-L-4-F (omis dans le tableau pour plus de clarté).
B. La protéine C10orf24 (Human_A, qui a été nommé STOX1 par la suite) a une structure secondaire similaire, caractéristique d'un domaine à « winged helix » comprenant la conservation de ces acides aminés hydrophobes. La règle Y/F spécifique pour les protéines FOX est absente dans C10orf24 et ses homologues, ce qui indique que ce gène représente un nouveau membre de la famille des protéines à « winged helix ». Le paralogue humain (DKFZp762K222) sur 4q35 est indiqué par Human_B. Le SNP identifié comme lié à la prééclampsie est indiqué en rouge : cette mutation (Y153H) implique le remplacement d'un acide aminé très conservé au cours de l'évolution comme en atteste cet alignement (Y ou F) et trouvé dans 46 des 48 femmes prééclamptiques.

Annexe 5 : Séquence en acides aminés des protéines chimériques 6Flag-STOX1A et 4Flag-STOX1B

Légende :

Séquence Flag

Séquence STOX1A ou STOX1B

DBD (*DNA Binding Domain*) : Site de fixation à l'ADN

6Flag-STOX1B

MDYKDDDDKGSDYKDDDDKGSEFMDYKDDDDKGSDYKDDDDKGSEFMDYKDDDDKG
SDYKDDDDKGSEFMARPVQLAPGSLALVLCRLEAQKAAGAAEEPGGRAVFRAFRRANAR
CFWNARLARAASRLAFQGWLRRGVLLVRAPPACLQVLRDAWRRRALRPPRGFRIRAVGD
VFPVQMNPITQSQFVPLGEVLCCAISDMNTAQIVVTQESLLERLMKHYPGIAIPSEDILYTTL
GTLIKERKIYHTGEGYFIVTPQTYFITNTTTQENKRMLPSDESRLMPASMTYLVSMESCAES
AQENAAPISHCQSCQCFRDMHTQDVQEAPVAAEVTRKSHRGLGESVSWVQNGAVSVSAE
HHICESTKPLPYTRDKEKGKKFGFSLLWRSLSRKEKPKTEHSSFSAQFPPEEWPVRDEDDLD
NIPRDVEHEIIKRINPILTVDNLIKHTVLMQKYEEQKKYNSQGTSTDMLTIGHKYPSKEGVK
KRQGLSAKPQGQGHSRRDRHKARNQGSEFQPGSIRLEKHPKLPATQPIPRIKSPNEMVGQK
PLGEITTVLGSHLIYKKRISNPFQGLSHRGSTISKGHKIQKTSDLKPSQTGPKEKPFQKPRSLD
SSRIFDGKAKEPYAEQPNDKMEAESIYINDPTVKPINDDFRGHLFSHPQQSMLQNDGKCCPF
MESMLRYEVYGGENEVIPEVLRKSHSHFDKLGETKQTPHSLPSRGASFSDRTPSACRLVDN
TIHQFQNLGLLDYPVGVNPLRQAARQDKDSEELLRKGFVQDAETTSLENEQLSNDDQALY
QNEVEDDDGACSSLYLEEDDISENDDLRQMLPGHSQYSFTGGSQGNHLGKQKVIERSLTEY
NSTMERVESQVLKRNECYKPTGLHATPGESQEPNLSAESCGLNSGAQFGFNYEEEPSVAKC
VQASAPADERIFDYYSARKASFEAEVIQDTIGDTGKKPASWSQSPQNQEMRKHFPQKFQLF
NTSHMPVLAQDVQYEHSHLEGTENHSMAGDSGIDSPRTQSLGSNNSVILDGLKRRQNFLQ
NVEGTKSSQPLTSNSLLPLTPVINV → 1 035 acides aminés

4Flag-STOX1B

MDYKDDDDKGSDYKDDDDKGSEFMDYKDDDDKGSDYKDDDDKGSEFMARPVQLAPGS
LALVLCRLEAQKAAGAAEEPGGRAVFRAFRRANARCFWNARLARAASRLAFQGWLRRG
VLLVRAPPACLQVLRDAWRRRALRPPRGFRIRAVGDVFPVQMNPITQSQFVPLGEVLCCAI
SDMNTAQIVVTQESLLERLMKHYPGIAIPSEDILYTTLGTLIKERKIYHTGEGYFIVTPQTYFI
TNTTTQENKRMLPSDESRLMPASMTYLDTESGI → 273 acides aminés.

Annexe 6 : Composition des gels et des tampons du Western blot

❖ Composition d'un gel à 10 % d'acrylamide :

Ingrédients	Gel de séparation (× 2)	Gel de concentration (× 2)
Acrylamide/bisacrylamide 30%	2,3 mL	1,3 mL
Tampon de séparation 4X	1,5 mL	-
Tampon de concentration 4X	-	2 mL
Glycérol 50%	1,2 mL	0,8 mL
Eau	5 mL	3,9 mL
APS 10 % (100 mg/mL)	60 µL	60 µL
Temed	10 µL	10 µL
<u>Remarque</u> : APS (= Ammonium Per Sulfate) et TEMED (=TEtraMethylEthyleneDiamine) sont deux réactifs qui vont permettre la polymérisation de l'acrylamide.		

Où :

Tampon de séparation 4X	Tampon de concentration 4X
Tris base 1,5 M	-
Tris HCl 0,5 M	Tris HCl 0,5 M
SDS 0,4%	SDS 0,4%

Annexe 7 : Composition d'un gel à 6 % d'acrylamide en condition non-dénaturante utilisé pour l'EMSA

Pour 10 mL (= 1 gel) :

- 2 mL du mélange 30% stock (29 % Acrylamide/1% Bis Acrylamide)
- 0,5 mL TBE 10X
- 7,5 mL H2O
- 100 µL APS 10 % (Ammonium PerSulfate)
- 10 µL TEMED

Annexe 8 : Séquences obtenues à l'issue de la PCR-sélection

Légende :

STRE1 (ou son complémentaire)

STRE2 (ou son complémentaire)

Chevauchement des deux séquences

Séquences obtenues avec la protéine 4Flag-STOX1B	Séquences obtenues avec la protéine STOX1B
>seq1 GTGGTGCGGGAACATGCTTCCACGGT >seq2 GACGCCATAAGTCTGTGTGCCCGGTT >seq3 CCACATGGCCCTAAGAGCAAGCATTT >seq4 CGGTGCGGAAATAATCATTTCACGG >seq5 TCGACTACGACATGGATTACATCTGAGCAACTGCAGCAAGGGCG >seq6 GGTGTTNCTGGAGGGAGGTGGTCATG >seq7 ACGCCGAAGGTTATGTTTAGTACTCA >seq8 TGCGCAGCGTTCTCGTCGGAAATTGC >seq9 GGTCTGGAATGGGTGCAAATTGCAGT (deux fois le motif 1) >seq10 CGGTGTGGAGAGCCATCTCACGGCAC >seq11 GTGGTGCGGGAACATGCTTCCACGGT >seq12 ATTTCGGACTTGCCGAGTTCCAGAGT >seq13 CGGTGTGGAGAGCCATCTCACGGCAC >seq14 GTCCGCTCTCTGTCTGTTGCCGTAGG >seq15 CTTTTGTAGCATGTATGTATGAGGGG >seq16 CTTTGGGTCCTGGACGTCCGTTGTCA >seq17 CTTTGGGTCCTGGACGTCCGTTGTCA >seq18 GCTTGACATTTCACGCGTTACCCTGG >seq19 ATGCAAAATCTGCCAAGATCCACCAA >seq20 ATGCAAAATCTGCCAAGATCCACCAA >seq21 ACAAAAAAACCAATACCGCCTACACC >seq22 CAGTTCGGAACCCAACGATAACATAC >seq23 ACCGTGGAAGCATGTTCCCGCGCCAC (les 2 motifs sur le brin complémentaire) >seq24 ATAGGTGCGGAGAGTTCATCACACGG >seq25 CGGTGTGGAGAGCCATCTCACGGCAC >seq26 TGCCGTTCGGGATTAGGGCAATCGNG >seq27 CGCGTTGGCAGGTTTGTCGCCACCAT >seq28 CGGTGGGCGCGCTTGACTTATCCTGT >seq29 CGGTGTGGAGAGCCATCTCACGGNAC >seq30	>seq1 GTTCCCTGTGATATTGTGGGATCGGT >seq2 AAGGTTGAGCCTATGTGTGTGCCATT >seq3 GGGGCGTTAGTCCAATTCAGCCTTGG >seq4 GGCCTTCCGTTCAGCCAGATAGGTAG >seq5 ACTGTAAGCGCATTCCGGATTATTTG >seq6 GAGAGTCGCTGTTTATGAAGAGTGAT >seq7 GTGACGTTTCAATTTGGGTATTGGTC >seq8 TGCCCTGGACTTAGAGTCGTTAGTTG >seq9 ATCGGTATTGTAATATGCTGTGAGAC >seq10 GCGGTCGGGATCGTGTAGATTGATCC >seq11 GAAGTGCATGAGTATGCTGGATGGTA >seq12 TGAGGGTGGCGGTAGACTTGACTTCA >seq13 TGTGGTAGAACTGCGGATGCGGGTTC >seq14 GCGGGGATGAGCGATGGTCTGCTCTG >seq15 CGTGTTGGGCGGGTTTGGCATAGC >seq16 AGCTTGCTGATTTGTGCTGATCAATT >seq17 TCAGAGGTCATCAGCTCGAGCGAGC >seq18 TCAACAGCCACCCCCAAAAACTTAAC >seq19 CACCTTAACCCACCCTTGATGTAACA >seq20 GGAGCGTGATACCTCAACCGCCCATC >seq21 AAACACATTTAATACCACTGTAAAAT >seq22 TATAGGAAAGCACAGGTGATCCAACA >seq23 GGTAGCGTTTCTGTTCAAGTCGCCAG >seq24 CTCTGATTATGCTTGTTATGTCGGGG >seq25 GTACTGGTTGAAGAGTATAAGTGGTC >seq26 CGCTAGTGTAGAGGAGCGGTACGCCA >seq27 TGTGGGTTCGTGCGGGTGGGTGTGCT >seq28 GAGCATCGAGGGGGGCTATTTGTTCT >seq29 GACGTCTTTTGCTTTATGTTTTCTAC >seq30 TAAGCGGGGTATAAATTAGGCTGAGG >seq31

ATAGTANTGGCGCANGTATGCGGGTG
>seq31
GAATAGTTAGTGTTCGTTTGCNTCNN
>seq32
CGGTGTGGAGAGCCATCTCACGGCAC
>seq33
TGAACCTACAGGGTGCGGTGCATGAT
>seq34
GGTGCGGATTGGTCAACTCACGGACT (deux fois le motif 1 et une fois le motif 2)
>seq35
CTACCCCACACGGGTCAATACAGATC
>seq36
CTCCCCGCNACACACCAGTCCACGCA (motif 1 sur le complémentaire, motif 2 sur ce brin)
>seq37
GCGCTCTCNGGNGCATTGGAGGTGCA
>seq38
GTGGTGCGGGAACATGCTTCCACGGA
>seq39
AGATTGGTGGCACATGCCAGGTGGAA
>seq40
ATTTTGGATTTTGAGTATACTCTGCC
>seq41
ACGAGGAGCCATGCCTTGGGCGTCCAG (sur le brin complémentaire)
>seq42
AATTGGGCATCTTAATGGTTTTAACT
>seq43
TGGGAACTTGTTAGCGCTGGCCAAGT
>seq44
ATGTCAATGGTCCCCTGTGTTATGGT
>seq45
GTTAGCAGTGTACGTTGACATCTAT
>seq46
GAAGAGATCAAATGTTGGATTTTTGT
>seq47
GTCTGCTATGTTCCGTTTGTAGAAAT
>seq48
GTTGCCAAGAATTTTAATGTGGCTCT
>seq49
AACCGTGAATTCGACAATTCCGCACC (les 2 motifs sur le brin complémentaire)
>seq50
GTGCCGTGAGATGGCTCTCCACACCG (les 2 motifs sur le brin complémentaire)
>seq51
AGTAGACACACAACAGCCATTATATA
>seq52
TCTGACTGCGAAGAACTTTCTCGTTT
>seq53
ATGGTGGCGACAAACCTGCCAACGCG
>seq54
CGTGAAAAGGTAAATTTTCCGCCCCCA (les 2 motifs sur le brin complémentaire, motif 2 tronqué)
>seq55
ACCGTGGAAGCATGTTCCCGCACCAC (les 2 motifs sur le brin complémentaire)
>seq56
GTGCCGTGAGATGGCTCTCCACACCG (les 2 motifs sur le brin complémentaire)
>seq57
AGGCCGCTTTCCGCCGGGGGCGCACG (motif 2 tronqué)
>seq58
TTTGGTCGCGTTGGGTTTCCAAGATG
GTAGCCTTGTGATGGCGAGGATCGTA
>seq32
GCCTCTGCCCATTGGGCTGGTTTTGG
>seq33
CGCAGGTGAGTACAATCGTCGTTCGT
>seq34
GTCAAAGCGTCAGATATGACATTTTT
>seq35
CGGTTCGTTAATCTCTTGTGGCATCT
>seq36
TGCGTGGNCTAGTCATTTAGGCCTTG
>seq37
AGTGTGGGCAATATATTTATTTAAGG
>seq38
TCAATAGCAACGCGAACGAGTGGCCC
>seq39
TTGTGAAGATAGCTTACCCGATTCN
>seq40
GGTTTATTGGAGATAATTATCCGGGA
>seq41
CTGGTACTCCCATTCCTACCCTAGAA
>seq42
TATCGGGATTTGGTTTTGGGTGCGGT
>seq43
GGGGGGTTACAACGAAAGGGGCGTTC
>seq44
ATGTAGGCTTGTCAATCTACGAATTT

LES FACTEURS DE TRANSCRIPTION CHEZ LES MAMMIFÈRES, MALADIES ASSOCIÉES, MÉTHODES D'ÉTUDE ET APPLICATION DE LA PCR-SÉLECTION SUR STOX1

Aurélien DUCAT

Résumé

Les facteurs de transcription sont classiquement définis comme des protéines ayant la capacité de se lier à des séquences spécifiques d'ADN et de réguler la transcription des gènes. Les facteurs de transcription ont longtemps fasciné les biologistes moléculaires de par leurs actions dans la régulation complexe de l'expression génique et leur importance dans les processus biologiques tels le métabolisme cellulaire ou la différenciation des cellules au cours du développement embryonnaire. Ainsi, des mutations au niveau de ces protéines entraînent souvent des répercussions importantes sur l'organisme, et sont notamment la cause de plusieurs maladies héréditaires potentiellement létales chez les mammifères. Au cours de mon stage de recherche dans le laboratoire du docteur Daniel Vaiman à l'Institut Cochin, j'ai travaillé sur un facteur de transcription récemment découvert : STOX1 (*Storkhead box 1*). Un polymorphisme de cette protéine a été associé à une maladie humaine de la grossesse, potentiellement létale : la prééclampsie. Mon travail a consisté à identifier le site de fixation à l'ADN de la protéine STOX1 en essayant de déterminer une séquence consensus via la méthode de PCR-sélection. Le principe de cette technique est de purifier, au sein d'une banque d'oligonucléotides aléatoires, des oligonucléotides sur lesquels un facteur de transcription se fixe. L'analyse bio-informatique m'a permis de mettre en évidence deux séquences consensus majoritaires que j'ai commencé à valider par deux méthodes : du retard sur gel et des essais luciférase. Les études poussées de ces protéines permettent d'entrevoir les premières applications scientifiques comme l'induction de cellules souches pluripotentes.

Mots clés

GÉNÉTIQUE / MALADIE HÉRÉDITAIRE / ADN / FACTEUR DE TRANSCRIPTION / STOX1 / STORKHEAD BOX 1 / PCR-SÉLECTION / MAMMIFÈRE

Jury

Président : Pr.
Directeur : M^{me} Marie ABITBOL
Assesseur : M. Laurent TIRET
Invité : M^{me} Fanny PILOT-STORCK

TRANSCRIPTION FACTORS IN MAMMALS, ASSOCIATED DISEASES, METHODS AND APPLICATION OF PCR-SELECTION ON STOX1

Aurélien Ducat

Summary

Transcription factors are proteins typically defined as having the ability to bind to specific DNA sequences and regulate gene transcription. Transcription factors have long fascinated molecular biologists through their actions in the complex regulation of gene expression and their importance in biological processes such as cellular metabolism and cell differentiation during embryonic development. Thus, mutations in these proteins often have a large impact on the organism, and include the cause of several potentially lethal hereditary diseases in mammals. During my research training in the laboratory of Dr Daniel Vaiman at the Cochin Institute, I worked on a transcription factor recently discovered: STOX1 (*Storkhead box 1*). A polymorphism of this protein has been associated with a human disease of pregnancy, potentially lethal: preeclampsia. My project aimed at identifying the DNA binding site of STOX1 protein, by trying to determine a consensus sequence using the PCR-selection method. The principle of this technique is to purify oligonucleotides on which a transcription factor binds, in a library of random oligonucleotides. The bioinformatics analysis allowed me to highlight two major consensus sequences that I started to validate by two methods: the gel shift assay and the luciferase assay. Extensive studies of these proteins allow a glimpse of the first scientific applications such as induction of pluripotent stem cells.

Keywords

GENETIC / HEREDITARY DISEASE / ADN / TRANSCRIPTION FACTOR / STOX1 / STORKHEAD BOX 1 / PCR-SELECTION / MAMMAL

Jury

President: Pr.
Director: Mrs Marie ABITBOL
Assessor: Mr Laurent TIRET
Guest: Mrs Fanny PILOT-STORCK