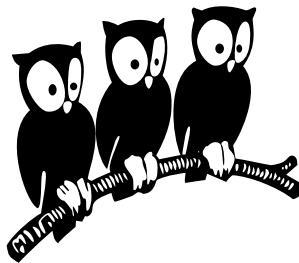


Département de Physique
École Normale Supérieure

Laboratoire de Physique Statistique



THÈSE de DOCTORAT de l'UNIVERSITÉ PARIS 7

Spécialité : Physique Théorique

présentée par

Marc SANTOLINI

pour obtenir le grade de DOCTEUR de l'UNIVERSITÉ PARIS 7

**Analyse computationnelle des éléments cis-régulateurs
dans les génomes d'eucaryotes supérieurs**

Soutenue le ZZ septembre 2013
devant le jury composé de :

- | | |
|------------------------------------|--------------------|
| M. Vincent HAKIM | Directeur de thèse |
| M. Martin Weigt | Rapporteur |
| M. Emmanuel Barillot | Examinateur |
| M. Alain Zider | Président du jury |
| M. Massimo Vergassola | Rapporteur |
| M. Pascal Maire | Membre invité |

Remerciements

...

Table des matières

Liste des figures	vii
Principales abréviations utilisées	ix
Avant-propos	1
Chapitre 1 - Introduction générale.	3
1.1 Le phénotype cellulaire	5
1.2 Les réseaux de régulation génétique	10
1.3 Les interactions protéine-ADN : modèles mathématiques	20
1.4 Les interactions protéine-ADN : mesures expérimentales	27
1.5 Les modules de cis-régulation (CRMs)	37
1.6 Prédition et validation des CRMs	50
1.7 Bases de données	59
Chapitre 2 - Modèles de fixation des Facteurs de Transcription à l'ADN.	67
2.1 Les modèles de fixation	69
2.2 Description des données biologiques	70
2.3 Présentation de l'algorithme	71
2.4 Performance des modèles	72
2.5 Analyse des corrélations	72
2.6 Comparaison avec des données <i>in vitro</i>	72
Chapitre 3 - <i>Imogene</i> : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle	75
3.1	77
Chapitre 4 - Étude de la différenciation épidermale chez la drosophile	79
4.1	81
Chapitre 5 - Étude de la différenciation musculaire chez la souris	83
5.1	85
Conclusion	87
Résumé	105

Liste des figures

Introduction générale.	3
1.1 Le paysage de la différenciation cellulaire	6
1.2 Spécification spatio-temporelle du type cellulaire	8
1.3 Différents exemples de reprogrammation cellulaire	9
1.4 Vision cybernétique du traitement de l'information par la cellule	11
1.5 Un réseau de régulation génétique type	13
1.6 Caractéristiques de l'épigénome	15
1.7 Exemples de motifs dans les réseaux de régulation génétique	16
1.8 Évolution du réseau transcriptionnel : l'exemple de la régulation myogénique. .	18
1.9 Différents états du facteur de transcription	20
1.10 Construction et utilisation du modèle PWM	23
1.11 Étapes d'une expérience de ChIP-on-chip et ChIP-seq	33
1.12 Résolution des expériences ChIP-on-chip et ChIP-seq	35
1.13 Expérience d'empreinte à la DNase I chez la levure : vers une résolution au nucléotide près	36
1.14 Les différents types de CRMs et leurs marques épigénétiques	38
1.15 Différents <i>enhancers</i> conduisent à différents patterns d'expression	40
1.16 Deux modèles d' <i>enhancers</i> : enhanceosome et billboard	42
1.17 L'enhanceosome de l'interferon- β	43
1.18 Flexibilité du code de cis-régulation au cours de l'évolution chez les <i>Drosophiles</i> .	44
1.19 Évolution de la fixation de HNF4 α chez les mammifères	46
1.20 « Shadow enhancer » du gène de segmentation <i>Hunchback</i>	48
1.21 De l' <i>enhancer</i> au super- <i>enhancer</i>	49
1.22 Différentes approches pour la prédiction des CRMs	51
1.23 Méthodes de validation des CRMs par transfection et transgenèse	56
1.24 Impact physiologique de la délétion et de la mutation d'un enhancer	57
1.25 Évolution du coût de séquençage	59
1.26 Distribution des tailles intergéniques et introniques chez différentes espèces .	61

Liste des figures

1.27 Visualisation de données ChIP-seq <i>via</i> le site UCSC	63
1.28 Les différentes données obtenues par le projet ENCODE	64
Modèles de fixation des Facteurs de Transcription à l'ADN.	67
2.1 Description graphique de l'algorithme.	71
Imogene : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle	75
Étude de la différenciation épidermale chez la drosophile	79
Étude de la différenciation musculaire chez la souris	83

Principales abréviations utilisées

ARNm	ARN messager
bHLH	<i>basic Helix-Loop-Helix</i>
bp	Paire de base
ChIP	Immunoprécipitation de la chromatine (<i>Chromatin immunoprecipitation</i>)
CRM	Module de cis-régulation (<i>Cis-Regulatory Module</i>)
DHS	Hypersensible à la DNase I (<i>DNaseI-hypersensitive</i>)
ESC	Cellule souche embryonnaire (<i>Embryonic Stem Cell</i>)
ISH	Hybridation <i>in situ</i> (<i>In-Situ Hybridization</i>)
kb	kilobases (1000bp)
MRF	Facteur de régulation myogénique (<i>Myogenic Regulatory Factor</i>)
nt	Nucléotide
PCR	Réaction en chaîne par polymérase (<i>Polymerase Chain Reaction</i>)
PWM	Matrice de poids (<i>Position Weight Matrix</i>)
TF	Facteur de transcription (<i>Transcription Factor</i>)
TFBS	Site de fixation d'un facteur de transcription (<i>Transcription Factor Binding Site</i>)
TSS	Site d'initiation de la transcription (<i>Transcription Start Site</i>)

Avant-propos

Cette thèse se présente sous la forme suivante...

Voici quelques remarques sur la version pdf de ce manuscrit, qui peuvent rendre la lecture plus aisée. Dans la table des matières, la liste des figures et la liste des annexes, les titres sont des liens hypertexte qui pointent vers l'item décrit. Dans la liste des notations utilisées et la bibliographie, ce sont les numéros de page qui sont des liens hypertexte.

these : version du lundi 1^{er} juillet 2013 à 16 h 57

Avant-propos

Chapitre 1

Introduction générale.

1.1 Le phénotype cellulaire	5
1.1.1 Qu'est-ce que le phénotype d'une cellule ?	5
1.1.2 La différenciation cellulaire	6
1.1.3 La cellule dans l'organisme : une spécification spatio-temporelle	8
1.1.4 La reprogrammation cellulaire	9
1.2 Les réseaux de régulation génétique	10
1.2.1 Vision cybernétique de la cellule	10
1.2.2 Divers modes de régulation	10
1.2.3 Câblage du réseau et fonction	17
1.2.4 Évolution des réseaux génétiques	17
1.3 Les interactions protéine-ADN : modèles mathématiques	20
1.3.1 Modes de recherche du site de fixation par le TF	21
1.3.2 Modèle PWM	21
1.3.3 Modèle biophysique	24
1.3.4 Modèle thermodynamique	25
1.4 Les interactions protéine-ADN : mesures expérimentales	27
1.4.1 Approches <i>in vitro</i> : MITOMI, SPR, PBM, CSI, SELEX, et HT-SELEX	28
1.4.2 Approche clonale : la technique de simple hybride	31
1.4.3 Approches <i>in vivo</i> : ChIP-on-chip, ChIP-seq, DNase I	32
1.5 Les modules de cis-régulation (CRMs)	37
1.5.1 Les différents types de CRMs	37
1.5.2 Grammaire des enhancers : enhanceosome vs billboard	41
1.5.3 Évolution des enhancers	43
1.5.4 Les « shadow enhancers »	47
1.5.5 Par delà les enhancers : les « super-enhancers »	48
1.6 Prédiction et validation des CRMs	50

Chapitre 1. Introduction générale.

1.6.1	Méthodes utilisant la concentration en sites de fixation	50
1.6.2	Méthodes utilisant la phylogénie	52
1.6.3	Méthodes utilisant les marques épigénétiques et de ChIP-seq pour des TFs	54
1.6.4	Validation expérimentale	55
1.6.5	Implication des CRMs dans les maladies humaines	57
1.7	Bases de données	59
1.7.1	Obtention de données génomiques	59
1.7.2	Obtention de données sur les TFs	61
1.7.3	Outils de visualisation	62
1.7.4	Le projet ENCODE	62

1.1 Le phénotype cellulaire

1.1.1 Qu'est-ce que le phénotype d'une cellule ?

Les organismes vivants sont constitués de cellules de l'ordre de quelques microns, facilement observables à l'aide d'un simple microscope optique. Chaque cellule contient un certain nombre de constituants (gènes, protéines, métabolites...) enclos par une membrane. Il existe des organismes unicellulaires (bactérie, levure) et multicellulaires (mouche, souris, homme). Ce sont ces derniers auxquels nous nous intéressons dans cette thèse. Les cellules qui les constituent sont majoritairement eucaryotes, c'est-à-dire qu'elles possèdent un noyau renfermant le matériel génétique.¹

Bien que possédant toutes le même matériel génétique, les cellules d'un organisme apparaissent d'emblée comme hétérogènes, que ce soit dans leur forme ou dans leurs constituants. Par exemple, chez l'homme, les érythrocytes ou globules rouges présents dans le sang sont des cellules de la forme d'un disque biconcave, dépourvues de noyau et riches en hémoglobine, tandis que les fibres musculaires squelettiques sont de forme longue et tubulaire, possèdent plusieurs noyaux et expriment actine et myosine.

Cette diversité semble néanmoins limitée. Aussi, parmi les $\sim 6 \cdot 10^{13}$ cellules du corps humain, on peut distinguer ~ 320 différents types cellulaires (Brazma et al., 2001). Bien entendu, ce nombre dépend du seuil de similarité choisi : deux cellules d'un même type n'expriment pas *exactement* le même nombre de molécules. Classiquement, la classification d'un type cellulaire se base sur des propriétés morphologiques observables au microscope ou encore sur l'analyse de molécules présentes à la surface des cellules. Par ailleurs, différents types cellulaires sont associés à différentes fonctions : dans notre exemple la fixation et le transport de l'oxygène dans le cas des globules rouges, la contraction dans le cas des fibres musculaires.

Ces différentes propriétés observables caractérisent le *phénotype* cellulaire (étymologiquement « exhiber un type » en grec). Nous allons le voir, ce phénotype est le résultat de la modulation par des facteurs environnementaux de l'expression génétique qui conditionne le contenu en protéines de la cellule.

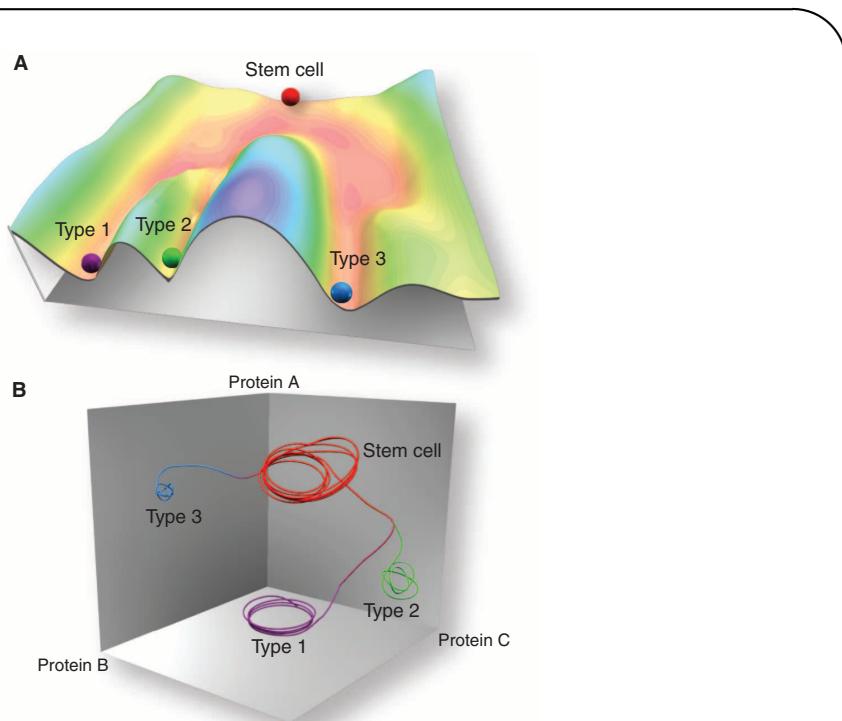


FIGURE 1.1 – Le paysage de la différenciation cellulaire.

Figure tirée de ([Furusawa and Kaneko, 2012](#)). A. Paysage épigénétique tel qu’imaginé par Waddington ([Waddington et al., 1957](#)) en résonance avec la notion de paysage énergétique en physique. Le développement cellulaire est représenté par une bille dévalant un paysage composé de différentes vallées séparées par des barrières difficilement franchissables, représentant les différents types cellulaires et leur robustesse face aux fluctuations. B. Représentation dynamique de l’évolution des états cellulaires. Le phénotype est ici caractérisé par l’expression de trois protéines A, B et C, dont l’évolution dans le temps peut être représentée par une trajectoire dans un espace tridimensionnel. Les états souches et différenciés sont caractérisés par des bassins d’attraction correspondant à différents types cellulaires.

1.1.2 La différenciation cellulaire

L’acquisition d’un phénotype cellulaire particulier au sein d’un organisme est le sujet de la biologie du développement. Cette acquisition passe par différentes étapes de différenciation cellulaire. Schématiquement, au cours du développement d’un organisme, des cellules non différenciées ou souches empruntent un chemin unidirectionnel de différenciation qui

1. Il existe cependant quelques cas connus d’organismes multicellulaires procaryotes, dont les cellules ne possèdent pas de noyau, par exemple chez les bactéries magnétotactiques ([Keim et al., 2004](#)).

restreint peu à peu le nombre de types cellulaires qu'elles peuvent potentiellement devenir, passant de l'état souche totipotent à des états pluripotents successifs avant la différenciation finale. Ainsi, la formation des cellules somatiques, qui sont les cellules d'un organisme n'étant ni souches ni germinales (les cellules qui donnent naissance aux gamètes ou cellules sexuelles), est le résultat d'un processus de différenciation initial lors du développement embryonnaire au cours duquel les cellules souches issues de l'œuf donnent naissance à trois couches de tissus distinctes : l'endoderme (feuillet interne), l'ectoderme (feuillet externe) et le mésoderme (feuillet intermédiaire). Des différenciations successives ont ensuite lieu au sein de ces couches pour former divers organes tels que le tube digestif (endoderme), les muscles et les os (mésoderme), ou encore la peau et le système nerveux (ectoderme).

Dans un écrit aujourd'hui célèbre datant de 1957 ([Waddington et al., 1957](#)), Waddington proposa une représentation de ces différentes étapes sous la forme d'un paysage épigénétique semblable aux paysages énergétiques dont sont coutumiers les physiciens (fig 1.1A). Dans cette représentation, le processus de différenciation cellulaire est comparé à une bille dévalant une pente et dont la trajectoire suit les multiples embranchements de vallées escarpées, chacune représentant un état de développement différent. Les vallées sont séparées par des pics dont la hauteur reflète la difficulté de passer d'un état à un autre, et les destinations finales possibles de la bille correspondent aux différents types cellulaires.

La notion de trajectoire de différenciation peut être rendue plus parlante en adoptant une représentation de système dynamique. Comme nous l'avons vu en 1.1.1, la cellule contient de nombreux composants : gènes, protéines ou autres métabolites, qui pris dans leur ensemble déterminent à un instant donné l'état cellulaire. Il est ainsi possible de représenter l'état cellulaire à un temps donné comme un point dans un espace de grande dimension dans lequel chaque axe représente l'abondance d'un certain composant (fig 1.1B). De par leur rôle primordial dans la définition de l'état cellulaire, l'expression des protéines (et donc des gènes qui les produisent) domine généralement ces composants, et on parle de « niveau d'expression génétique » pour décrire leur abondance. Les changements d'expression génétique, au cours desquels certains gènes vont être activés et d'autres réprimés, induisent un changement de l'état cellulaire, ce qui se traduit par une trajectoire dans l'espace des états. Ces changements d'expression restreignent finalement l'état cellulaire à une certaine région, définie comme un « attracteur » de la dynamique. Une fois au sein d'un attracteur, l'état cellulaire est robuste aux perturbations du niveau d'expression génétique des différentes composantes. Les attracteurs

Chapitre 1. Introduction générale.

peuvent alors être vu comme des types cellulaires distincts correspondant aux différentes vallées de la représentation de Waddington ([Kaufmann, 1993](#)).

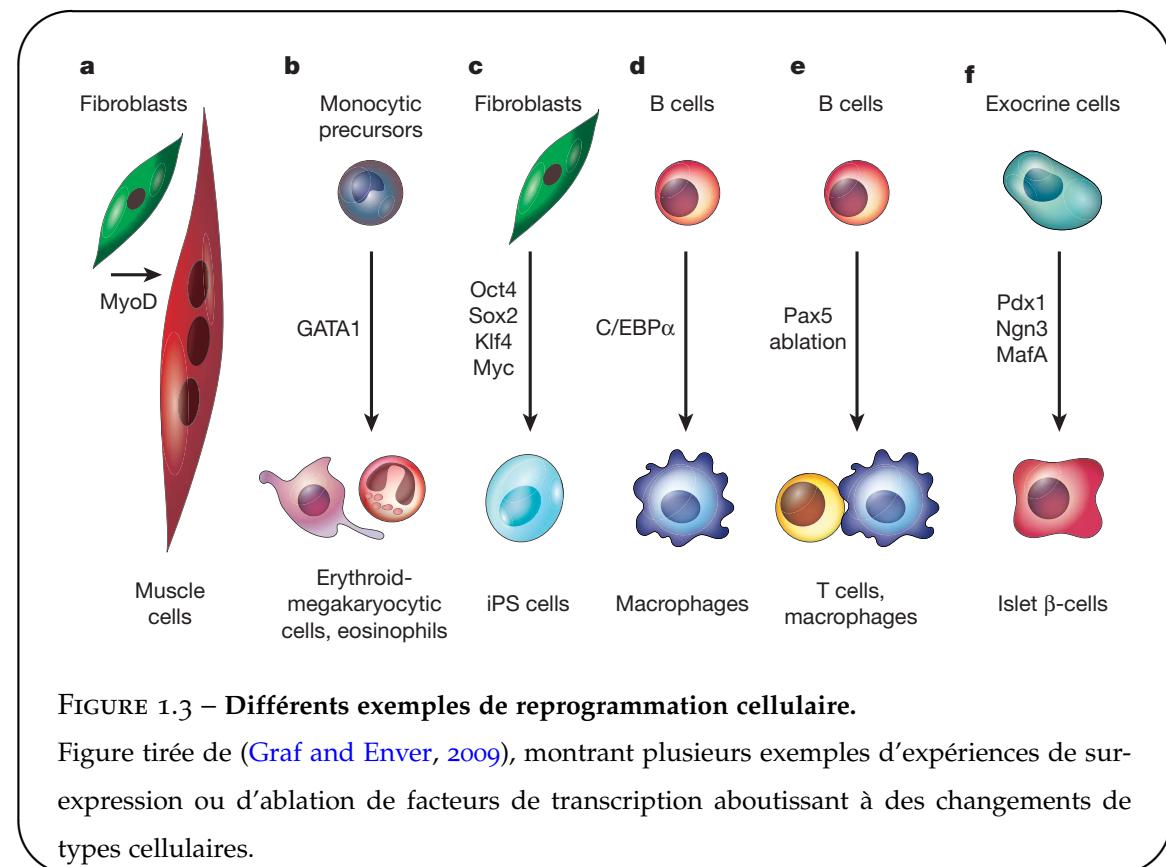
1.1.3 La cellule dans l'organisme : une spécification spatio-temporelle



FIGURE 1.2 – Spécification spatio-temporelle du type cellulaire.

Hybridation *in situ* de l'ARN du gène *Myog*, marqueur de la différenciation des progéniteurs du muscle squelettique, chez des embryons de souris âgés de 9.5, 10.5 et 11.5 jours (de gauche à droite), observés sous un même grossissement de 10. Le motif (*pattern*) de spécification du muscle squelettique est clairement visible au niveau des somites, les futures vertèbres. Images tirées de la base de donnée Embrys (<http://embrys.jp>).

Au sein de l'organisme, la différenciation cellulaire opère à un rythme précis et dans un contexte cellulaire bien défini. Aussi, les trajectoires dans l'espace d'expression génétique que nous avons présentées précédemment sont fonction de l'espace – la position de la cellule dans l'organisme, qui détermine en particulier la concentration des signaux qu'elle reçoit de son environnement – et du temps – le stade de développement de l'organisme -. Il est ainsi possible d'observer chez l'embryon certains motifs ou *patterns* spatio-temporels d'expression génétique correspondant à des organes en formation et révélés par hybridation *in situ* de l'ARN de certains gènes spécifiques d'un type cellulaire. Par exemple, dans le cas de la formation des muscles squelettiques, le gène de différenciation terminale *Myog* est exprimé chez la souris dès 8 jours embryonnaires au niveau des somites, segments correspondant aux futures vertèbres de la souris adulte, puis commence à être exprimé au niveau des bourgeons de membres à 11.5 jours (voir fig 1.2).



1.1.4 La reprogrammation cellulaire

Depuis plusieurs décennies, différentes expériences ont exhibé la plasticité des états différenciés, élargissant ainsi considérablement la vision classique selon laquelle des cellules souches totipotentes se différencient de manière irréversible en des cellules de moins en moins plastiques, jusqu’à atteindre un état différencié stable. Par exemple, ([Blau et al., 1985](#)) ont montré en 1985 que des programmes d’expression génétique dormants peuvent être exprimés de manière dominante dans des cellules différencierées par la fusion de différents types cellulaires : ainsi, la fusion de cellules musculaires avec des cellules non musculaires permettait l’activation des gènes de type musculaire dans le type cellulaire non musculaire. Puis différents travaux ont montré qu’il était possible de convertir des lignées de cellules différencierées en un autre type cellulaire en introduisant certaines protéines régulatrices de la transcription, ou Facteurs de Transcription (TFs) ([Davis et al., 1987; Kulessa et al., 1995](#)) : on parle alors de trans-différenciation, dont l’un des exemples canoniques est la différenciation de cellules de la peau ou fibroblastes en cellules musculaires par l’introduction du facteur de différenciation myogénique MyoD (voir fig 1.3). Parallèlement, des expériences réalisées chez plusieurs espèces

Chapitre 1. Introduction générale.

de mammifères ont montré que le transfert de noyaux de cellules différencierées embryonnaires ou adultes dans un oeuf énucléé peut mener à la formation d'un organisme complet, montrant de manière univoque que l'identité des cellules différencierées peut être complètement renversée ([Gurdon and Melton, 2008](#)). Enfin, l'avancée la plus récente dans ce domaine a été la démonstration que des cellules somatiques différencierées peuvent être reprogrammées en cellules souches puripotentes par simple introduction d'un « cocktail » de 4 facteurs de transcription : Oct4, Sox2, c-Myc et Klf4 ([Takahashi and Yamanaka, 2006](#)) (fig 1.3C).

1.2 Les réseaux de régulation génétique

Afin de pouvoir mieux comprendre les mécanismes de différenciation et de reprogrammation exposés en 1.1, il convient de se plonger dans les mécanismes internes de la cellule qui régissent ses changements d'états.

1.2.1 Vision cybernétique de la cellule

Le paradigme qui règne sur la biologie moléculaire depuis plus d'un demi siècle est celui des réseaux génétiques. L'expression des gènes est en effet régulée par des protéines, les facteurs de transcription, qui sont eux-mêmes issus de l'expression d'autres gènes, créant ainsi un réseau d'interactions entre gènes. Certaines protéines peuvent par ailleurs directement réguler l'activité d'autres protéines, et certains ARNs issus de la transcription de gènes non codants jouent aussi un rôle fondamental dans la régulation de l'activité génétique, le tout formant un réseau complexe d'interactions à différents niveaux. La compréhension de ce réseau et des fonctions qui en résultent forme le socle de la biologie des systèmes. Dans ce cadre, la cellule est vue comme une unité de traitement d'information qui interprète différents signaux reçus en entrée, les traite par un réseau interne de régulation, et réagit en sortie en modifiant son état ou son comportement (fig 1.4). L'intérêt d'une telle description mécanistique est qu'elle permet d'opérer quantifications mathématiques et prédictions, ce qui l'a rendue extrêmement fertile au cours des dernières décennies ([Nurse and Hayles, 2011](#)).

1.2.2 Divers modes de régulation

Les modes de régulation qui permettent à la cellule d'interpréter des signaux afin de changer d'état sont nombreux. Nous allons nous concentrer ici sur ceux affectant la production

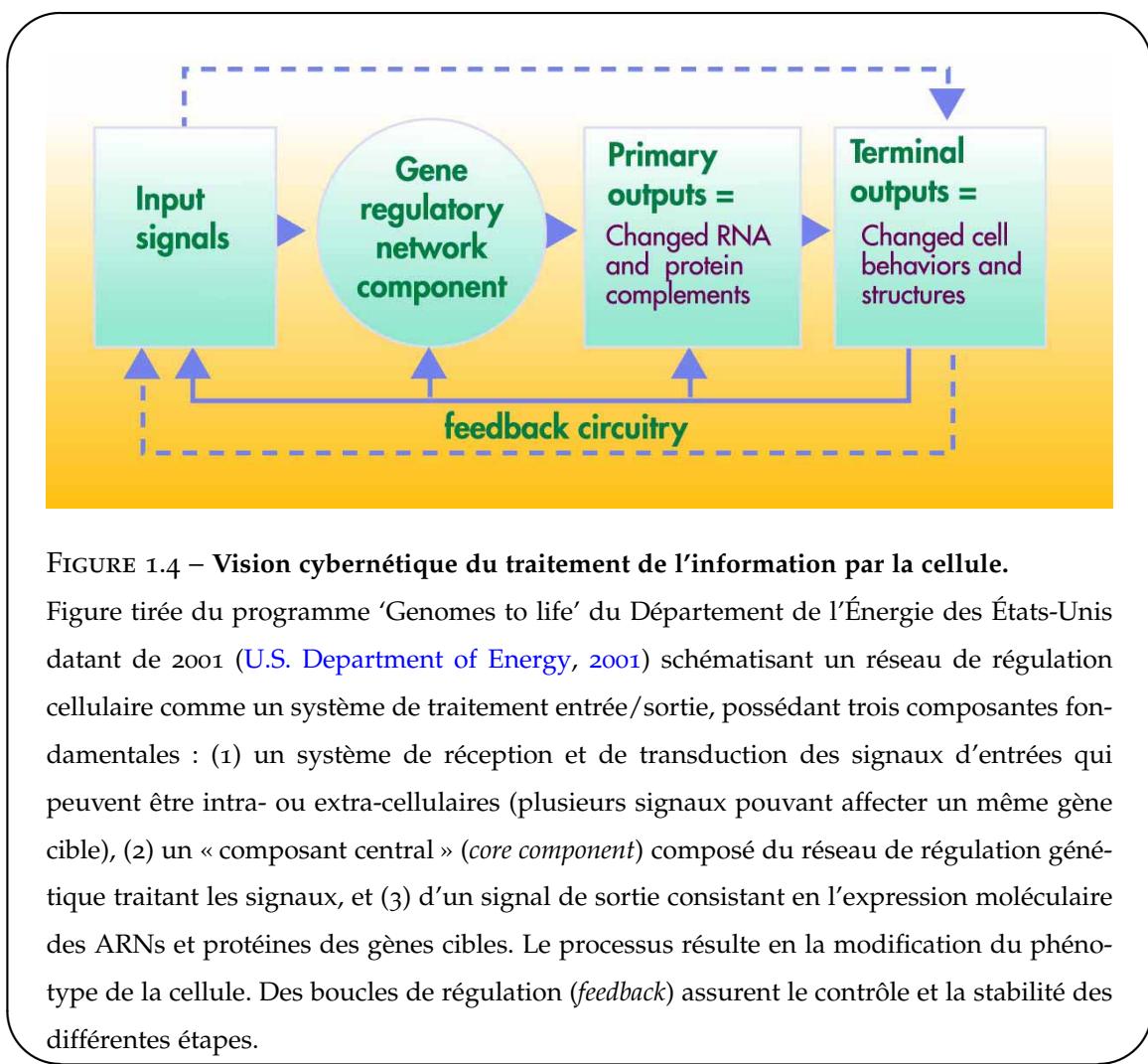


FIGURE 1.4 – Vision cybernétique du traitement de l’information par la cellule.

Figure tirée du programme ‘Genomes to life’ du Département de l’Énergie des États-Unis datant de 2001 ([U.S. Department of Energy, 2001](#)) schématisant un réseau de régulation cellulaire comme un système de traitement entrée/sortie, possédant trois composantes fondamentales : (1) un système de réception et de transduction des signaux d’entrées qui peuvent être intra- ou extra-cellulaires (plusieurs signaux pouvant affecter un même gène cible), (2) un « composant central » (*core component*) composé du réseau de régulation génétique traitant les signaux, et (3) d’un signal de sortie consistant en l’expression moléculaire des ARNs et protéines des gènes cibles. Le processus résulte en la modification du phénotype de la cellule. Des boucles de régulation (*feedback*) assurent le contrôle et la stabilité des différentes étapes.

Chapitre 1. Introduction générale.

d'ARNs ou de protéines (fig. 1.5).

- **Régulation génétique**

Tout d'abord, un réseau d'expression génétique est caractérisé par un jeu d'interactions entre différents gènes. Ces interactions se font par l'intermédiaire de protéines régulatrices appelées facteurs de transcription ou TFs, qui sont au nombre de ~ 1400 chez l'homme ([Vaquerizas et al., 2009](#)), soit 6% des protéines encodées. Les gènes qui les expriment représentent donc ~ 3% de l'ensemble des 30,000 gènes connus à ce jour. Pour réguler (activer ou inhiber) la transcription d'un gène cible, les TFs se fixent sur des sites de reconnaissance spécifiques sur l'ADN de ~ 10bp et interagissent avec la machinerie transcriptionnelle au niveau du promoteur du gène cible. Les TFs peuvent se fixer sur le promoteur même, comme c'est souvent le cas chez la bactérie, ou dans des régions distales allant jusqu'à plusieurs centaines de kb, comme on trouve plus couramment chez les organismes complexes. Par ailleurs, différents TFs peuvent se combiner sur certaines régions de régulation contenant de multiples sites de fixation pour former des complexes protéiques. Ces régions, appelées modules de cis-régulation (CRMs) ou plus communément *enhancers*, sont d'une taille typique de ~ 1000bp et ont la particularité de conduire à une expression spatio-temporelle très spécifique du gène cible. Ces différents points seront amplement développés en section 1.5.

- **Régulation épigénétique**

Outre la régulation génétique, due à l'action de protéines issues de séquences codantes et se fixant sur des séquences d'ADN – régulation qui est donc entièrement encodée dans le génome et transmise à la descendance –, il existe un autre mode de régulation de la transcription des gènes qui permet notamment d'acquérir une modification d'expression génétique transmise à la descendance sans qu'il y ait modification du code génétique : c'est la régulation épigénétique. Cette régulation passe notamment par la modification des propriétés chimiques de l'ADN et des histones sur lequel il s'enroule pour former la chromatine (fig. 1.6). Ainsi, la méthylation des dimères CpG de l'ADN² au niveau des régions riches en CG, ou îlots CpG, situées près de nombreux promoteurs et habituellement dépourvues de ces marques conduit à une inactivation du gène cible ([Bird, 2002](#)). Par ailleurs, la méthylation des histones au niveau des résidus lysines entraîne la fermeture de la chromatine, empêchant l'expression du ou des

². Les dimères C-G sont appelés CpG, où p caractérise le phosphore liant les deux bases, pour les différencier du CG utilisé pour parler de la statistique en C et G de l'ADN

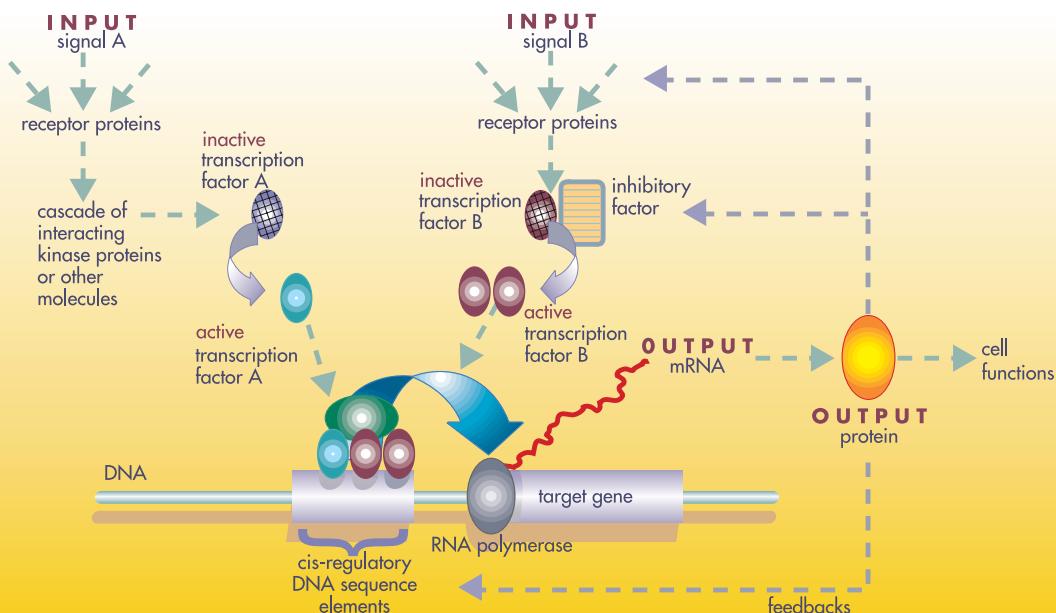


FIGURE 1.5 – Un réseau de régulation génétique type.

Dans cette représentation schématique tirée du rapport du U.S. Department of Energy (2001), deux voies de signalisation A et B transmettent des signaux d'entrée (qui peuvent être intra ou extra cellulaires) en rendant des facteurs de transcription actifs. Une fois activés, ces derniers interagissent avec des séquences d'ADN proches d'un gène cible en se fixant sur des sites de petite taille ($\sim 10\text{bp}$). Les différents facteurs de transcription interagissent entre eux pour former des complexes occupant des régions de $\sim 1000\text{bp}$ appelées modules de cis-régulation ou CRMs (voir section 1.5). Lorsque les facteurs de transcription sont fixés sur le CRM de leur gène cible, il peuvent activer ou inhiber la transcription d'ARN et donc la production de la protéine correspondante.

Chapitre 1. Introduction générale.

gène(s) situés à leur niveau, alors que l'acétylation des mêmes lysines entraîne au contraire une ouverture de la chromatine, favorisant ainsi la transcription génétique ([Greer and Shi, 2012](#)). Ce mode de régulation sera développé plus en détails en section [1.5.1](#).

• Régulation post-transcriptionnelle

Les modifications post-transcriptionnelles affectent les ARNs issus de la transcription des gènes. Ces modifications peuvent être causées par des microARNs ou miRNAs qui sont des ARNs de ~ 23 nts issus d'ARNs se repliant en structure double brin de type « épingles à cheveux » ou *hairpins*. Les miRNAs s'associent à la protéine *Argonaute* du complexe RISC (*RNA-induced silencing complex*) pour entraîner la dégradation spécifique d'ARNms ([Bartel, 2009](#)). De manière similaire, certains *hairpins* de taille plus importante sont clivés par la protéine Dicer pour former plusieurs petits ARNs de taille similaire aux miRNAs : ce sont les siRNAs (*small interfering RNAs*). Ceux-ci recrutent aussi le complexe protéique RISC et ciblent spécifiquement des ARNm ([Hammond et al., 2001](#); [Hannon, 2002](#)). Ce phénomène est connu sous le nom d'interférence ARN (RNAi) et a donné lieu à une méthode aujourd'hui couramment utilisée pour inhiber l'expression d'un gène.

• Régulation post-traductionnelle

Les modifications post-traductionnelles affectent les protéines issues de la traduction des ARNs. Elles passent par une modification chimique des protéines, typiquement la phosphorylation, ou comme nous l'avons vu pour la régulation épigénétique, la méthylation ou l'acétylation. Ces modifications peuvent avoir pour effet de changer l'activité de la protéine, que ce soit en modifiant son activité enzymatique ou en déclenchant sa relocalisation nucléaire. Il existe aussi des modifications de structure de la protéine, comme c'est le cas du facteur de transcription *Shavenbaby* chez la Drosophile : dans sa forme native, cette protéine inhibe la transcription de ses gènes cible ; cependant ses résidus terminaux peuvent être clivés par des petits peptides de 11 à 32 acides aminés encodés par le gène *Pri*, rendant la protéine transcriptionnellement active ([Kondo et al., 2010](#)).

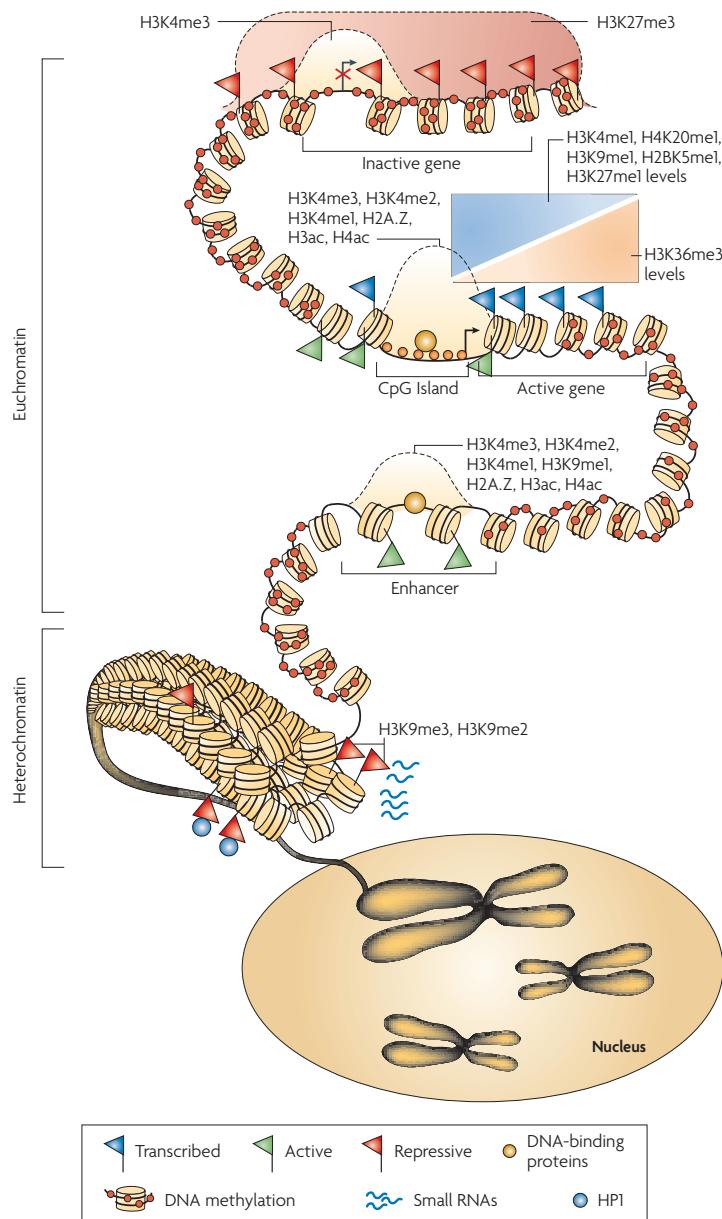


FIGURE 1.6 – Caractéristiques de l'épigénome.

Figure tirée de ([Schones and Zhao, 2008](#)). Les chromosomes sont partagés entre régions accessibles d'euchromatine et régions difficilement accessibles d'hétérochromatine. Les régions hétérochromatiques sont marquées par la di- et triméthylation de la lysine 9 de l'histone H3 (H3K9me2 et H3K9me3). La méthylation de l'ADN est répandue à travers tout le génome, mais est absente de certaines régions comme les îlots CpG, les promoteurs et les CRMs. La modification H3K27me3 couvre de larges régions englobant des gènes inactifs. Les marques H3K4me3, H3K4me2, H3K4me1 et l'acétylation des histones marquent les TSSs des gènes actifs. Les marques H3K4, H3K9, H3K27, H4K20 et H2BK5 marquent les régions transcris activement à proximité de la région 5' des gènes (en amont), alors que la marque H3K36 marque les gènes transcrits dans leur région 3' (en aval).

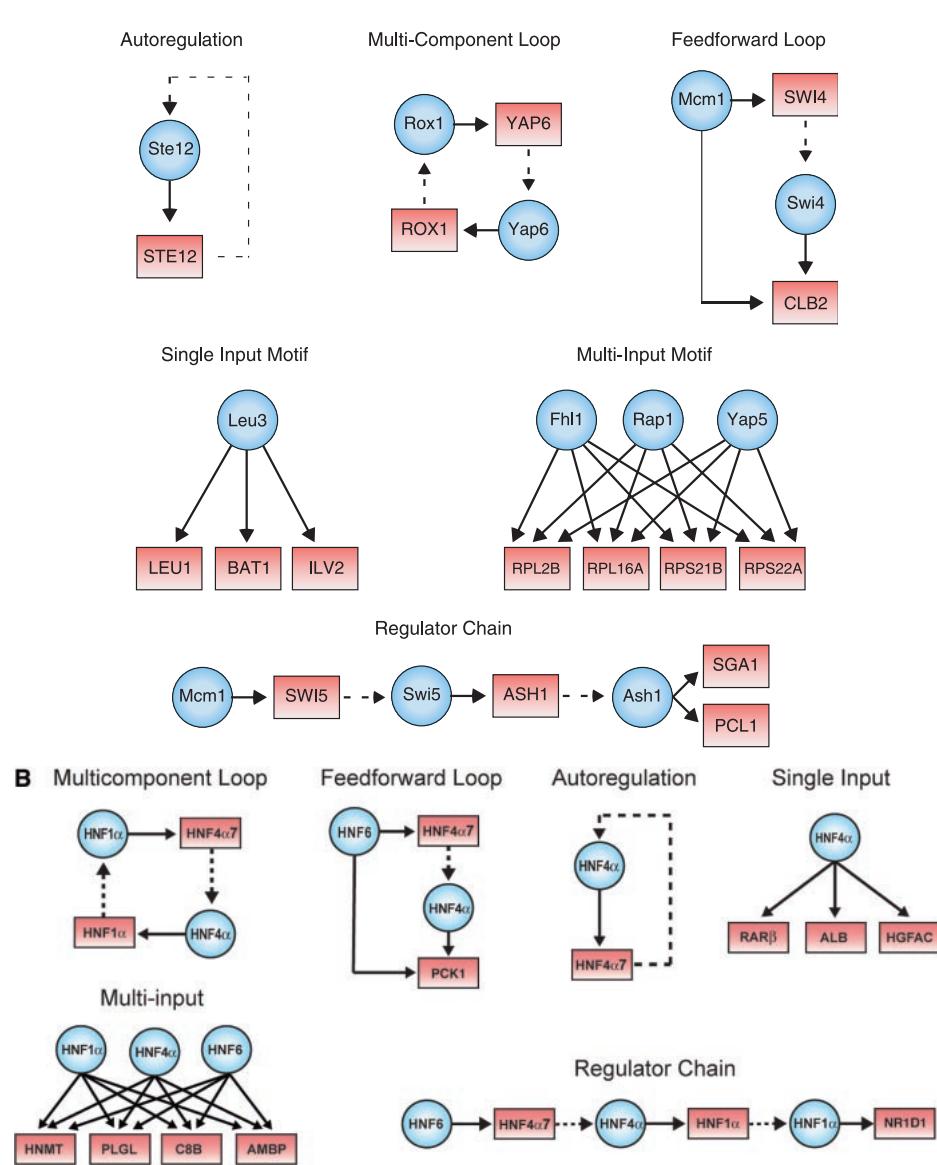


FIGURE 1.7 – Exemples de motifs dans les réseaux de régulation génétique.

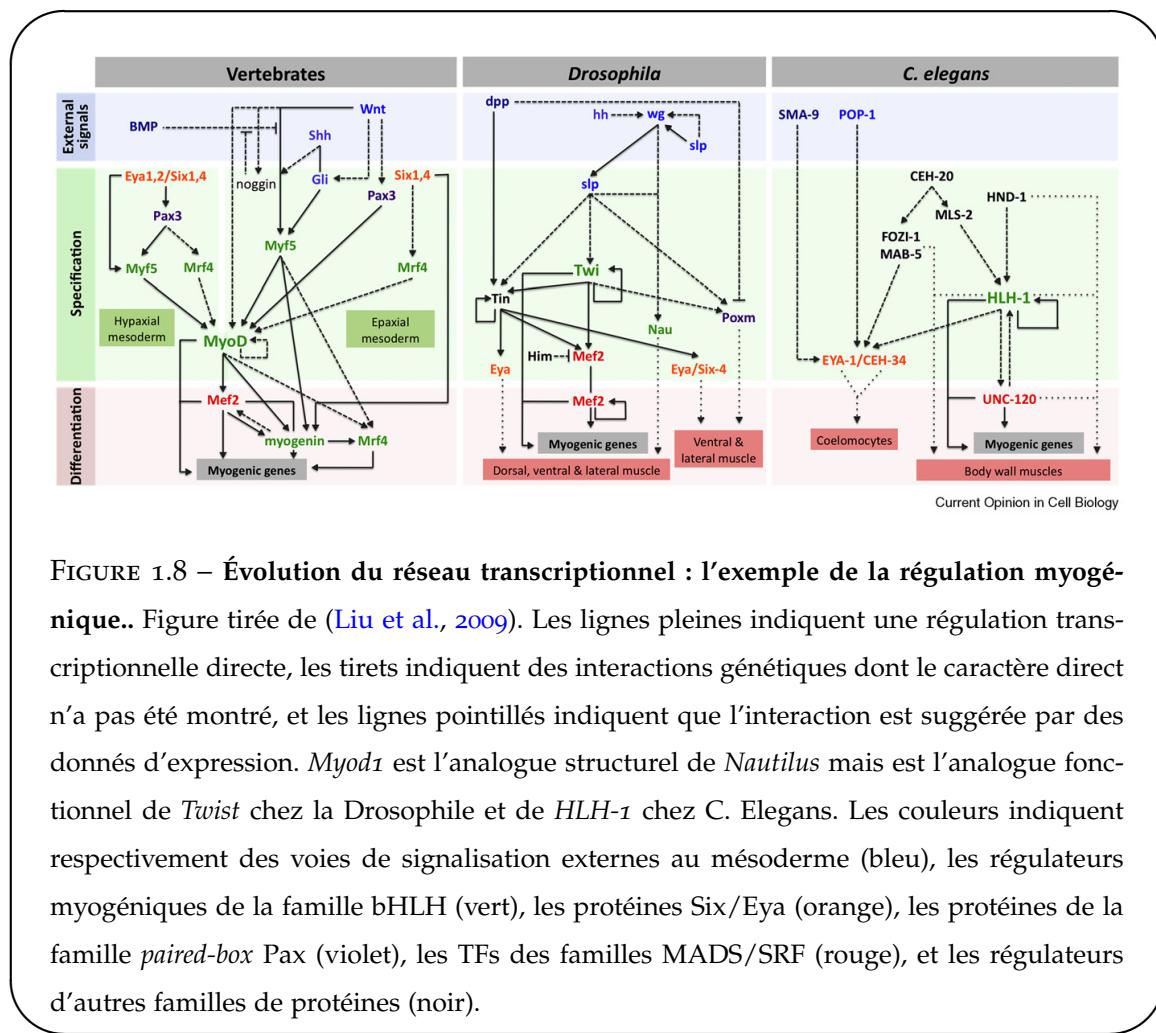
Ces exemples sont issus d'analyses d'interactions entre facteurs de transcription (cercles bleus) et promoteurs (rectangles rouges) (A) chez la levure (Lee et al., 2002) et (B) chez l'homme (Odom et al., 2004). Les flèches solides indiquent la fixation d'un facteur de transcription à un promoteur, et les flèches en pointillé indiquent l'expression d'un facteur de transcription à partir de son gène.

1.2.3 Câblage du réseau et fonction

Maintenant que nous avons vu la nature des interactions au sein des réseaux génétiques, nous pouvons nous pencher sur leur structure. Notamment, plusieurs études réalisées chez divers organismes de la bactérie à l'homme ont révélé que les réseaux de transcription contiennent un petit ensemble de motifs de régulation récurrents, appelés motifs de réseaux (Alon, 2007a; Shen-Orr et al., 2002; Milo et al., 2002) (fig. 1.7). Ces motifs peuvent être vus comme les pièces élémentaires servant à la construction de réseaux fonctionnels. De tels motifs furent d'abord détectés de manière systématique chez la bactérie *Escherichia coli* en remarquant qu'ils apparaissaient dans le réseau de transcription bien plus souvent qu'on ne l'attendrait dans un réseau aléatoire (Shen-Orr et al., 2002). Les mêmes motifs ont ensuite été trouvés chez la levure (Milo et al., 2002; Lee et al., 2002) et chez l'homme (Odom et al., 2004). La récurrence de ces motifs est liée aux fonctions qu'ils remplissent. Par exemple, la boucle d'autorégulation négative, qui est trouvée chez la moitié des répresseurs d'*Escherichia coli*, possède deux fonctions : l'une est de parvenir rapidement à un état d'équilibre en utilisant un promoteur fort, l'autre est de servir de tampon au bruit d'expression (Alon, 2007b). Un autre motif récurrent est la boucle feedforward. Celle-ci consiste en 3 gènes : un régulateur X, qui régule Y, tous deux régulant Z. Dans le cas où des interactions sont des activations et que X et Y sont requis pour activer Z, cette boucle peut servir de tampon au bruit d'expression de X, évitant que des fluctuations de son niveau d'expression n'entraîne par erreur l'activation de Z.

1.2.4 Évolution des réseaux génétiques

L'importance des motifs est rendue plus claire lorsque l'on s'intéresse à l'évolution des réseaux. En effet, au cours de l'évolution, les réseaux de régulation génétique changent : modification des constituants, recâblage du réseau, duplication d'éléments... Néanmoins, certaines modifications sont plus défavorisées du point de vue évolutif que des autres. Ainsi, les motifs tels que les boucles d'autorégulation ou les boucles feedforward, de par leur importance fonctionnelle, auront tendance à être conservés. Pour ce qui est des éléments du réseau, la modification d'un régulateur, par exemple la mutation d'un acide aminé au sein d'un facteur de transcription, aura des conséquences sur l'ensemble des éléments régulés par ce facteur de transcription et pourra donc être fortement délétère. Par contre, la modification d'un site de reconnaissance de ce facteur de transcription sur l'ADN n'aura qu'une portée locale sur la



régulation du gène associé.

À titre d'exemple, prenons le cas du réseau de différenciation du muscle squelettique présenté en figure 1.8, que nous étudierons plus en détail dans le chapitre 5 de ce manuscrit. Au cœur de ce réseau génétique se trouvent les facteurs de régulation myogéniques ou MRFs, des facteurs de transcription de type bHLH qui ont la capacité de convertir des cellules non mesodermiques, c'est-à-dire n'étant pas destinées à devenir des progéniteurs musculaires, en cellules ayant des propriétés musculaires (Weintraub et al., 1989). Ces facteurs sont dits « régulateurs maîtres » de la différenciation musculaire. Chez les vertébrés il y a quatre MRFs : *Myf5*, *Mrf4*, *Myod1*, qui ont des rôles redondants dans la spécification des progéniteurs musculaires, et *Myog*, qui conduit à la différenciation terminale. Chez la Drosophile c'est le TF *Twist* qui semble être le principal MRF, mais contrairement aux MRFs des vertébrés, son rôle ne s'arrête

pas au contrôle de la différenciation musculaire mais est plus général dans le développement du mésoderme (Baylies et al., 1998). C'est cependant le gène *Nautilus* qui possède la séquence d'acides aminés la plus proche de celle des MRFs vertébrés. Ce dernier permet la spécification des progéniteurs myogéniques, et son expression est restreinte au développement musculaire. Néanmoins, les mutants *nautilus* sont viables et son rôle semble mineur comparé aux MRFs vertébrés. Enfin, chez le ver *Caenorhabditis elegans*, c'est l'orthologue de *Myod1*, *hh-1*, qui tient rôle de MRF.

Malgré ces différences (nombre de MRFs, membre de la famille bHLH tenant ce rôle), on retrouve dans les trois cas une boucle feedforward conservée au niveau de la régulation des cibles des MRFs (fig. 1.8). Ainsi, MyoD régule l'expression de Mef2 et l'activité de MAPK p38 en même temps que l'expression de plusieurs cibles initiales, et par la suite MyoD et phospho-Mef2 co-régulent des gènes plus tardifs. Ce mécanisme permet ainsi de réguler l'aspect temporel de l'expression génétique. Chez la Drosophile, le même motif est observé avec Twist et Mef2 et chez C. Elegans avec HLH-1 et le TF UNC-129, de la même famille que Mef2.

Le cœur du réseau est donc conservé dans la forme (topologie), même s'il y a des divergences dans le fond (membres de la famille de TFs impliqués). Néanmoins, les éléments régulateurs en amont, ainsi que les membres périphériques du réseau ont rapidement évolué. Par exemple, chez les vertébrés le TF Pax3 est très en amont dans la hiérarchie génétique et permet l'activation des MRFs et la spécification myogénique, alors que chez la Drosophile son homologue *poxm* est en aval des MRFs et sa perte de fonction n'a que des effets mineurs sur la myogenèse. Par ailleurs, le complexe composé de protéine Six et de leur cofacteur Eya, initialement découvert comme régulateur majeur de la différenciation oculaire chez la Drosophile, est chez les vertébrés un régulateur essentiel situés en amont des MRFs. Chez la Drosophile, il possède aussi un rôle dans la spécification myogénique, mais bien plus en aval que chez les vertébrés. Enfin, chez C. Elegans ce complexe est aussi en aval des MRFs mais il participe en plus à la détermination de cellules non myogéniques.

Nous voyons donc que l'évolution d'un réseau génétique possède de multiples facettes : conservation de motifs de réseau fonctionnellement importants (dans notre exemple, la boucle feedforward au cœur du réseau régissant l'aspect temporel de l'expression des cibles), recâ-

Chapitre 1. Introduction générale.

blage des interactions pour traiter différents signaux d'entrée... Par ailleurs, il apparaît que plus qu'à des TFs particuliers, c'est à des familles de TFs que nous avons affaire. Aussi un même rôle au sein du réseau peut-il être rempli par différents membres d'une même famille, comme c'est le cas pour *Myod1* et *Twist*. Ceci s'explique par le fait que les membres d'une même famille partagent des propriétés d'interaction avec l'ADN semblables. Ces interactions sont à la source du fonctionnement du réseau, et nous allons maintenant présenter plus en avant leurs propriétés.

1.3 Les interactions protéine-ADN : modèles mathématiques

Nous l'avons vu, les interactions entre facteurs de transcription et ADN sont une composante essentielle des réseaux génétiques. Les TFs se fixent sur des sites spécifiques de ~ 10 bp dans le voisinage des gènes qu'ils régulent. Trouver ces sites est donc un premier pas vers la reconstruction des réseaux de régulation sous-jacents. Dans cette section nous présentons les modèles d'interactions protéine-ADN qui ont été proposés, et leur application concrète à la recherche de sites de fixation.

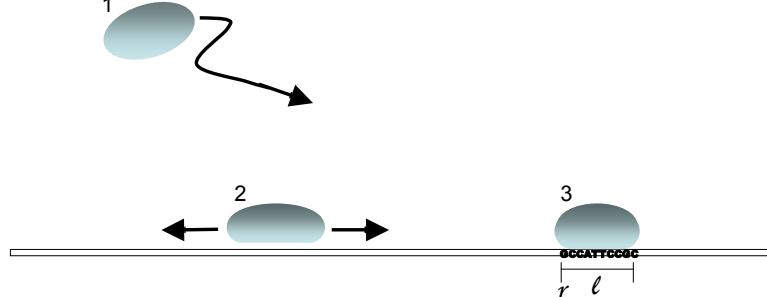


FIGURE 1.9 – Différents états du facteur de transcription. Figure tirée de ([Lässig, 2007](#)).

Lors de sa recherche de site de fixation, le TF peut se trouver dans trois états distincts : (1) un état libre de diffusion tridimensionnelle, (2) un état de diffusion unidimensionnelle sur l'ADN par fixation non spécifique, et (3) un état de fixation spécifique. L'énergie de fixation dépend du site de fixation, de taille l et de coordonnée r .

1.3.1 Modes de recherche du site de fixation par le TF

Un facteur de transcription peut être dans plusieurs états : en diffusion tridimensionnelle, auquel cas il est dit « libre », ou bien fixé sur l'ADN. Dans ce dernier cas, il interagit avec l'ADN selon deux modes : une attraction non spécifique d'énergie E_{ns} indépendante de la position sur l'ADN, et une interaction spécifique $E_s(r)$ qui dépend de la séquence de taille $l \sim 10$ à la position r sur l'ADN. L'interaction non spécifique est due à l'interaction électrostatique entre la protéine chargée positivement et l'ADN chargé négativement, alors que l'interaction spécifique implique des liaisons hydrogènes entre le domaine de fixation de la protéine et les nucléotides du site de fixation. La protéine passe d'un mode à l'autre en changeant de conformation. Le facteur de transcription peut ainsi se trouver dans trois états thermodynamiques représentés en figure 1.9 : en diffusion tridimensionnelle libre, fixé non spécifiquement (diffusion unidimensionnelle le long de la structure d'ADN), et fixé spécifiquement sur l'ADN. Ces trois modes contribuent à la cinétique de la recherche d'un site fonctionnel (Berg et al., 1981; Winter and von Hippel, 1981; Winter et al., 1981). Ainsi, l'attraction non spécifique conduit la protéine à passer à peu près autant de temps fixé sur l'ADN qu'en diffusion libre. La recherche de site de reconnaissance est donc un processus mixte de diffusion unidimensionnelle sur l'ADN et de diffusion tridimensionnelle dans le milieu. Lorsqu'il est fixé sur l'ADN, le facteur diffuse dans un paysage d'énergie E_{ns} plat lorsqu'il est dans sa conformation de fixation non spécifique, ou dans un paysage d'énergie $E_s(r)$ dans sa conformation de fixation spécifique. Cela permet au facteur d'échantillonner les sites de faible énergie $E_s(r)$ tout en évitant d'être bloqué par les barrières de haute énergie en passant en mode de recherche non spécifique. Ce processus s'avère au final très efficace : les temps de recherche sont typiquement inférieurs à une minute, ce qui est petit devant les processus de régulation de la cellule qui se déroulent au mieux sur quelques minutes (Gerland et al., 2002; Slutsky and Mirny, 2004). Il est donc pertinent de décrire l'effet d'un site de fixation sur la régulation d'un gène cible par la probabilité qu'il a de fixer un facteur de transcription à l'équilibre thermodynamique.

1.3.2 Modèle PWM

Présenté en 1987 par Berg et von Hippel (Berg and von Hippel, 1987), le modèle PWM est le modèle le plus simple décrivant l'énergie de fixation spécifique entre un facteur de transcription et un site de fixation sur l'ADN. Ce modèle repose sur plusieurs hypothèses.

Chapitre 1. Introduction générale.

Tout d'abord, il y a l'hypothèse importante que les sites de fixation des TFs sur l'ADN ont été sélectionnés au cours de l'évolution pour leur propriété de sites de reconnaissance, quelle que soit la concentration du TF dans la cellule. En d'autres termes, le processus de sélection discrimine les sites de fixation sur la seule base de leur énergie de fixation à un TF donné : les sites ayant une énergie de fixation dans une certaine gamme sont retenus, les autres rejetés. Par ailleurs, au sein de cette gamme d'énergie « utile », toutes les séquences sont équiprobables. Enfin, la dernière hypothèse est que chaque nucléotide d'un site de fixation contribue de manière indépendante, c'est-à-dire additive à l'énergie totale du site. Cette hypothèse permet de simplifier le problème en gardant le nombre de paramètres petit.

L'argument de Berg et von Hippel est que ce problème est analogue à celui de physique statistique consistant à déduire les taux d'occupation des niveaux d'énergie de particules indépendantes sachant que l'énergie totale doit avoir une certaine valeur moyenne E . La solution de ce problème est donnée par la formule de Boltzmann reliant énergie et taux d'occupation :

$$f_{i,b} = \exp(-\lambda E_{i,b}) / \mathcal{Z}_i \quad (1.1)$$

où $f_{i,b}$ est la probabilité d'observer la base b à la position i du site de fixation, $E_{i,b}$ est l'énergie associée (en $k_B T$), \mathcal{Z}_i est la fonction partition qui permet de normaliser la distribution à la position i , et λ est un facteur sans dimension, analogue du β de la thermodynamique, et lié au processus de sélection. Dans la suite, nous intégrerons ce facteur à l'énergie.

La connaissance des fréquences des bases permet de définir une autre quantité utile caractérisant la variabilité des séquences de fixation, l'information relative des sites par rapport à une séquence d'ADN aléatoire ([Stormo and Fields, 1998](#)) :

$$\mathcal{I} = \sum_{i=1}^L \sum_{b=A,C,G,T} f_{i,b} \ln \left(\frac{f_{i,b}}{\pi_b} \right) \quad (1.2)$$

où L est la taille du site de fixation et π_b correspond à la probabilité *a priori* d'observer la base b dans le génome. Parce que l'énergie est définie à une constante près, il est usuel de la définir relativement au fond génomique :

$$\tilde{E}_{i,b} = \ln \left(\frac{f_{i,b}}{\pi_b} \right) \quad (1.3)$$

L'énergie totale d'un site S_i est alors

1.3. Les interactions protéine-ADN : modèles mathématiques

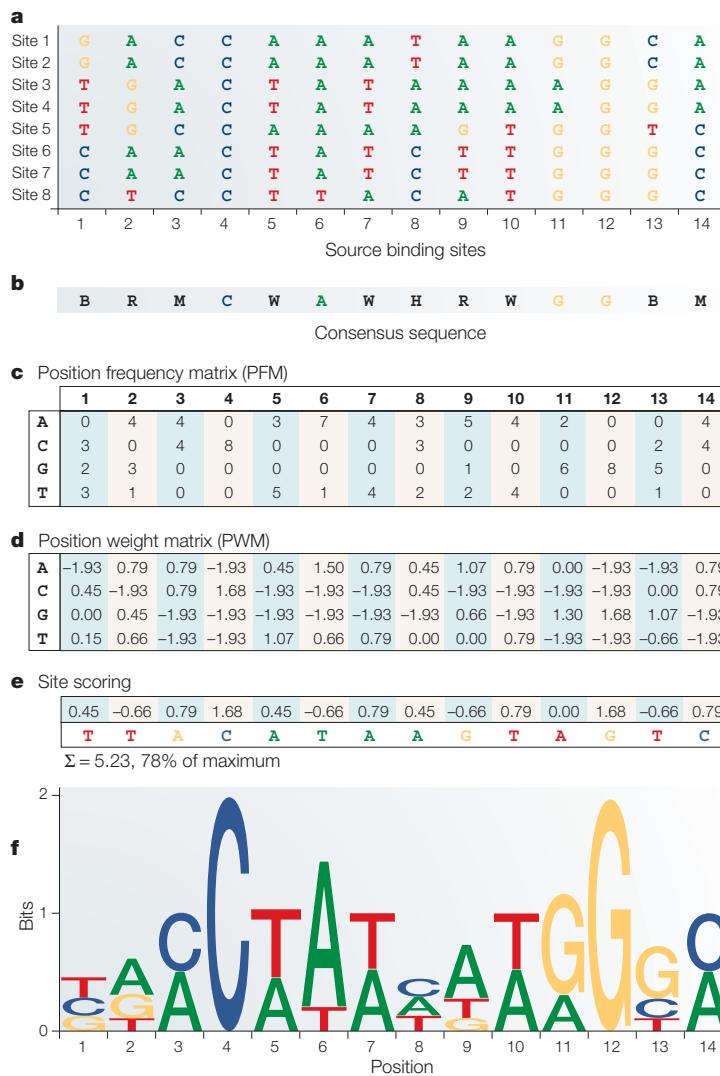


FIGURE 1.10 – Construction et utilisation du modèle PWM. Figure tirée de (Wasserman and Sandelin, 2004). (a) Supposons connus un certain nombre de sites de fixation d'un facteur de transcription (dans ce cas MEF2). (b) Séquence consensus correspondante utilisant les symboles IUPAC. (c) Une matrice de fréquence est construite, indiquant pour chaque nucléotide sa multiplicité à une position donnée dans le site. (d) La PWM est simplement construite en prenant le logarithme relatif des fréquences PWMs par rapport aux fréquences *a priori* des nucléotides. (e) Le score (ou énergie) d'une séquence d'ADN donnée est calculé en additionnant les poids PWM correspondants. (f) La PWM peut être représentée sous forme de logo (Giocomo et al., 2011). Dans cette représentation, la hauteur d'une colonne représente le contenu en information ou information relative moyenne d'une position, et la taille des bases reflète leur fréquence.

$$\begin{aligned}
E &= \sum_{i=1}^L \tilde{E}_{i,b} \\
&= \sum_{i=1}^L \ln \left(\frac{f_{b(i)}}{\pi_b} \right) \\
&= \ln \left(\frac{\prod_{i=1}^L f_{b(i)}}{\prod_{i=1}^L \pi_b} \right) \\
&= \ln \left(\frac{P(S_i|\text{TF})}{P(S_i|\text{fond génomique})} \right)
\end{aligned} \tag{1.4}$$

où $b(i)$ est la base située à la position i du site de fixation. Cette énergie quantifie simplement à quel point la séquence S_i est plus ($E > 0$) ou moins ($E < 0$) probablement un site de fixation (de probabilité $P(S_i|\text{TF})$) qu'un site tiré au hasard dans le génome (de probabilité $P(S_i|\text{fond génomique})$). On parle aussi de *score* de la séquence. L'information relative \mathcal{I} , qui est le score moyen des séquences fixées par le TF, peut alors être vue comme quantifiant à quel point l'ensemble des sites de fixation se distingue d'un ensemble de même taille de sites tirés au hasard.

Avec ces outils en main, il devient alors simple de bâtir un modèle PWM et de l'utiliser pour prédire des séquences fixées (fig. 1.10). Étant donnés des sites de fixation connus, il suffit d'évaluer la fréquence d'occurrence de chaque base à chaque position. La comparaison avec les probabilités génomiques *a priori* d'occurrence permet alors de bâtir une matrice de score, la PWM. Cette matrice peut alors être utilisée pour attribuer un score à une séquence d'ADN en additionnant les scores à chaque position. Finalement, les séquences ayant un score dépassant un certain seuil sont considérées comme des séquences de fixation.

1.3.3 Modèle biophysique

Le modèle PWM est basé sur une hypothèse forte, celle que les sites de fixation ont été sélectionnés sur la base de leur seule affinité ou énergie envers un TF. Néanmoins, à aucun moment n'intervient la concentration du TF dans la cellule, dont dépend pourtant la probabilité de fixation. C'est ce que tente de capturer le modèle biophysique (Gerland et al., 2002; Djordjevic et al., 2003; Zhao et al., 2009).

Considérons l'interaction entre un TF et une séquence d'ADN S_i :



où $TF : S_i$ dénote le complexe entre le TF et le site S_i . La constante d'équilibre de cette réaction s'écrit selon la loi d'action de masse :

$$K_i = \frac{[TF : S_i]}{[TF][S_i]} \quad (1.6)$$

Le site peut être dans deux états : occupé par le TF ou libre. Aussi, la probabilité que le TF soit fixé au site s'écrit simplement

$$P(\text{fixation}|S_i) = \frac{[TF : S_i]}{[TF : S_i] + [S_i]} = \frac{1}{1 + \frac{1}{K_i[TF]}} = \frac{1}{1 + e^{\beta(E_i - \mu)}} \quad (1.7)$$

où $E_i = -kT \ln(K_i)$ est l'énergie libre standard de fixation (souvent notée ΔG), $\mu = kT \ln[TF]$ est le potentiel chimique, k est la constante de Boltzmann, T la température et $\beta = 1/kT$. Ici nous avons considéré qu'il n'y avait qu'un seul site de fixation. De manière générale, le site est en compétition avec le fond génomique, ce qui ajoute une contribution à μ (voir section 1.3.4). À l'instar du modèle PWM, l'énergie E_i est généralement prise comme étant une fonction additive des énergies individuelles des différentes bases du site. Ainsi, lorsque le TF est à faible concentration ($\mu \rightarrow -\infty$), le modèle biophysique écrit en équation 1.7 se réduit au modèle PWM.

1.3.4 Modèle thermodynamique

La description biophysique peut être réécrite en termes thermodynamiques en utilisant des raisonnements simples sur le nombre d'états possibles et leur énergie (et donc poids de Boltzmann) associée. Nous adoptons ici l'approche de (Gerland et al., 2002). On pourra par ailleurs se référer à l'excellente revue (Lässig, 2007). Considérons le cas simple d'un seul facteur de transcription interagissant avec un génome de taille $L \gg 1$ ne contenant qu'un seul site fonctionnel, le reste de la séquence étant aléatoire. Nous l'avons vu, l'expérience montre que la protéine se fixe à l'ADN avec une probabilité 1/2. Lorsqu'elle est fixée, elle est à l'équilibre entre le mode spécifique et le mode non spécifique. Nous désirons savoir avec quelle probabilité elle est fixée de manière spécifique. La fonction de partition, énumérant tous les poids de Boltzmann associés aux différents états accessibles au TF fixé s'écrit :

$$\mathcal{Z} = \sum_{r=1}^L e^{-\beta E_s(r)} + L e^{-\beta E_{ns}} \quad (1.8)$$

Notons i la position du site fonctionnel. On peut écrire :

$$\begin{aligned} \mathcal{Z} &= e^{-\beta E_s(i)} + e^{-\beta E_{ns}} + \sum_{r \neq i} e^{-\beta E_s(r)} + (L - 1) e^{-\beta E_{ns}} \\ &\simeq e^{-\beta E_i} + \mathcal{Z}_0 \end{aligned} \quad (1.9)$$

où \mathcal{Z}_0 est la fonction de partition d'une séquence aléatoire, et nous avons introduit l'énergie E_i définie par

$$e^{-\beta E_i} = e^{-\beta E_s(i)} + e^{-\beta E_{ns}} \quad (1.10)$$

Dans le cas d'un site de reconnaissance, $E_{ns} \gg E_s(i)$ de sorte que $E_i \simeq E_s(i)$ ([Gerland et al., 2002](#)). La probabilité que le facteur soit fixé sur le site fonctionnel s'écrit finalement :

$$P(\text{fixation spécifique}|E_i) = \frac{e^{-\beta E_i}}{\mathcal{Z}} = \frac{1}{1 + e^{\beta(E_i - F_0)}} \quad (1.11)$$

où $F_0 = -kT \log \mathcal{Z}_0$ est l'énergie libre d'une séquence génomique aléatoire. On reconnaît une fonction de Fermi, avec un seuil d'énergie à F_0 : pour $E_i < F_0$, la protéine est essentiellement fixée de manière spécifique à son site de reconnaissance, alors que pour $E_i > F_0$, elle ne distingue plus le site du fond génomique et y est faiblement fixée.

Généralisons à présent au cas de plusieurs facteurs de transcription et sites de reconnaissance. Nous négligeons le recouvrement entre facteurs de transcription fixés sur des sites proches, qui poserait des problèmes stériques et corrèlerait les sites de fixation dans un certain voisinage (la présence d'un TF empêchant la présence d'un autre), et considérons que le nombre de TFs est grand devant le nombre de sites de reconnaissance pour éviter les problèmes de saturation : ainsi, le génome est composé de L séquences indépendantes, chacune pouvant être soit non occupée, soit occupée de manière non spécifique, soit occupée de manière spécifique. Notons μ le potentiel chimique du TF en solution. La fonction de partition totale est le produit des fonctions de partition des sites indépendants,

$$\mathcal{Z}(\mu) = \prod_{r=1}^L \mathcal{Z}(\mu, r) \quad (1.12)$$

 1.4. *Les interactions protéine-ADN : mesures expérimentales*

où la fonction de partition d'un site s'écrit :

$$\mathcal{Z}(\mu, r) = e^{-\beta\mu} + e^{-\beta E_s(r)} + e^{-\beta E_{ns}} \quad (1.13)$$

En utilisant à nouveau la définition de E_i en éq. 1.10, la probabilité de fixation d'un site à la position i s'écrit

$$P(\text{fixation spécifique}|E_i) = \frac{e^{-\beta E_i}}{\mathcal{Z}(\mu, i)} = \frac{1}{1 + e^{\beta(E_i - \mu)}} \quad (1.14)$$

La valeur de μ est liée à la fois au nombre de TFs ainsi qu'à la possibilité de se fixer dans le fond génomique. Elle est fixée implicitement par l'équation :

$$n = \sum_{r=1}^L \frac{1}{1 + e^{\beta(E_r - \mu)}} \quad (1.15)$$

qui signifie simplement que le nombre de TFs n dans le système est égal à la somme sur tous les sites de fixation possibles pondérée par la probabilité que le TF y soit fixé. Lorsque $\mu \rightarrow -\infty$ et que la fonction de Fermi peut être approximée par la loi de Boltzmann, l'équation peut s'inverser et l'on trouve (Aurell et al., 2007)

$$\mu = F_0 + kT \log n \quad (1.16)$$

où F_0 est l'énergie libre du fond génomique introduite en éq. 1.11. Ainsi, la prise en compte d'une multiplicité de TFs ajoute un facteur $kT \log n$ au seuil de la fonction de Fermi par rapport au cas d'un seul TF. Par ailleurs, cette approche thermodynamique nous a permis de généraliser le modèle biophysique simple introduit en section 1.3.3.

1.4 Les interactions protéine-ADN : mesures expérimentales

Ces dernières années, des avancées technologiques considérables ont permis d'une part d'établir des modèles de fixation spécifique pour de nombreux TFs, d'autre part de localiser leurs sites de fixation dans le génome. Ces avancées ont eu lieu autant sur le plan *in vitro*, utilisant protéines purifiées et séquences nucléiques artificielles pour déduire l'affinité protéine-ADN, que sur le plan *in vivo*, mesurant l'interaction de la protéine avec l'ADN génomique (Stormo and Zhao, 2010).

1.4.1 Approches *in vitro* : MITOMI, SPR, PBM, CSI, SELEX, et HT-SELEX

- **Approche microfluidique : MITOMI**

En 2007, Maerkel et Quake ont mis au point une technique appelée MITOMI (Mechanically Induced Trapping Of Molecular Interactions) permettant une mesure directe de l'affinité d'un TF à des centaines de séquences d'ADN à la fois ([Maerkel and Quake, 2007](#)). Cette technique repose sur l'utilisation d'un système microfluidique composé de chambres dans lesquelles un fluide dont on peut facilement modifier la composition circule dans des canaux d'un diamètre de l'ordre de $1\mu\text{m}$ dont le microenvironnement est finement contrôlé. Le fluide contient des gènes synthétiques codant pour le TF ainsi que du matériel permettant la synthèse de la protéine directement au sein de la chambre, ce qui évite de purifier préalablement le TF. Chaque chambre du système contient des anticorps fixés à la surface permettant de capturer le TF et une certaine concentration d'une séquence d'ADN spécifique contenant une marque fluorescente. Le système contient ainsi des centaines de séquences d'ADN différentes, chacune étant présente à différentes concentrations. Lorsque le TF est fixé par les anticorps, il recrute des séquences d'ADN selon leur affinité. Celles qui ne se fixent pas sont lavées. Au final, les séquences fixées produisent un signal de fluorescence. La comparaison des signaux pour différentes concentrations d'ADN donne accès au rapport des constantes d'équilibre K_{eq} (eq. 1.6). La comparaison avec une séquence référence dont la constante K_{eq} est connue permet alors de déterminer le K_{eq} absolu pour chaque séquence de fixation.

En utilisant 17 systèmes de ce type, ils ont ainsi pu mesurer l'affinité de 4 TFs de type bHLH à 464 séquences d'ADN différentes : les séquences consensus et des séquences ayant une, deux, trois ou quatre mutations. À titre de comparaison, ils ont construit une PWM à partir des séquences contenant une seule mutation, puis ont prédit les énergies attendues des séquences à plusieurs mutations. La prédiction de la PWM s'est avérée bonne dans seulement 56% des cas pour les séquences à deux mutations, 10% pour les séquences à 3 mutations et 0% des cas pour les séquences à 4 mutations, montrant les limites de ce modèle indépendant confronté à des données d'interactions d'ordre supérieur. Un modèle plus raffiné prenant en compte l'énergie d'interaction non spécifique et incluant des interactions entre nucléotides voisins permet néanmoins de rendre compte des valeurs observées ([Stormo and Zhao, 2007](#)). Nous reviendrons sur la nécessité de prendre en compte les interactions entre paires de nu-

 1.4. *Les interactions protéine-ADN : mesures expérimentales*

cléotides lors de l’interaction spécifique entre TF et ADN dans le chapitre 2.

- **Approche physique : la microscopie SPR**

La méthode de résonance des plasmons de surface (*Surface Plasmon Resonance* ou SPR) est habituellement utilisée pour étudier l’interaction d’une protéine avec un ligand (qui peut être une autre protéine), mais elle peut aussi être utilisée pour mesurer les interactions entre une protéine et quelques centaines de séquences d’ADN différentes ([Shumaker-Parry et al., 2004](#); [Campbell and Kim, 2007](#)). Le principe de la microscopie SPR est que l’angle de réflexion de la lumière sur une fine surface d’or, par exemple, dépend de la masse de molécules fixées de l’autre côté de sa surface. Si de l’ADN est lié à la surface, la fixation du TF induit un changement de masse et donc d’angle de reflection lumineuse mesurable au cours du temps. Ainsi, la cinétique de fixation du TF jusqu’à l’atteinte de l’équilibre est accessible. On peut de même étudier la dissociation du TF lors du lavage de la surface. Ces mesures donnent directement accès aux taux d’association k_{on} et de dissociation k_{off} que la simple mesure de la constante d’équilibre $K_{eq} = k_{on}/k_{off}$ ne permet habituellement pas de déterminer.

- **Approches basées sur des puces à ADN : PBM et CSI**

L’analyse de fixation des protéines par puce à ADN (*Protein-Binding Microarray* ou PBM) est une technologie haut débit qui a été développée au cours des 10 dernières années ([Berger et al., 2006](#)). Les puces sont composées de 44,000 puits auxquels sont liés des brins d’ADN. Une puce contient tous les sites de fixation de 8bp possibles ($4^8/2 = 32,768$ séquences en prenant en compte le fait qu’il y a un site sur chacun des deux brins d’ADN) plus deux bases flanquantes (une à chaque extrémité) qu’il est possible de faire varier. Un TF purifié à partir de cellules ou synthétisé *in vitro* est ajouté à la puce, qui est ensuite lavée pour se débarrasser des fixations non spécifiques. La quantité de protéine fixée à un puits donné est déterminée grâce à un anticorps fluorescent contre la protéine. L’enrichissement en protéine est calculé relativement au bruit de fond (anticorps non spécifique par exemple). Il est alors possible d’utiliser ces mesures pour bâtir une PWM du TF (voir par exemple [Kinney et al. \(2007\)](#)).

Une autre méthode utilise aussi des puces à ADN : c’est l’identification de site apparenté (*Cognate Site Identifier* ou CSI) ([Warren et al., 2006](#)). Une différence technique avec les PBMs est que l’ADN est d’abord synthétisé en simple brin puis se replie en double brin pour former le site de fixation, évitant ainsi de devoir générer l’ADN double brin à partir de précurseurs.

Chapitre 1. Introduction générale.

Par ailleurs, le TF est en compétition avec un marqueur fluorescent qui peut se fixer à l'ADN : il n'est donc pas nécessaire d'utiliser un marquage spécifique sur le TF ou sur un anticorps, ce qui rend la procédure plus généralisable. Finalement, la spécificité du TF est représentée par un « paysage de spécificité » qui encapsule l'information de fluorescence de l'ensemble des variations par rapport à une séquence consensus dans une représentation simple ([Carlson et al., 2010](#)).

- **Approche par purification des séquences fixées : SELEX et HT-SELEX**

Mise au point il y a plus de 20 ans, la méthode SELEX (*Systematic Evolution of Ligands by EXponential enrichment*) repose sur la sélection de séquences d'ADN aléatoires par un TF *in vitro* ([Oliphant et al., 1989](#); [Tuerk and Gold, 1990](#); [Blackwell and Weintraub, 1990](#); [Wright et al., 1991](#)). Une bibliothèque de sites de fixation potentiels est d'abord générée en synthétisant des séquences d'ADN aléatoires ou en utilisant des séquences génomiques. Les extrémités de ces séquences contiennent des précurseurs permettant l'amplification exponentielle par PCR. Le TF purifié est ajouté aux sites et les séquences fixées sont séparées des séquences non fixées, par exemple par retard sur gel. Après un cycle de sélection, les séquences récupérées sont encore enrichies en séquences de basse affinité pour le TF, car celles-ci sont simplement initialement bien plus abondantes que les séquences de haute affinité. Afin d'augmenter la proportion de séquence de grande affinité, les séquences filtrées sont amplifiées puis filtrées à nouveau, ceci sur plusieurs cycles. À la fin de ce processus, les séquences sélectionnées sont clonées et séquencées, résultant en un nombre typique de moins de ∼ 100 séquences indépendantes ([Fields et al., 1997](#)). Si les séquences initiales sont issues d'ADN génomique, il est possible d'utiliser l'hybridation des séquences à des puces à ADN. La présence de plusieurs cycles de sélection rend néanmoins la détermination des énergies de fixation moins directe qu'avec les techniques précédentes. Une variante de la technique appelée SELEX-SAGE utilise des multimères de sites à la place de sites uniques et permet de réduire le nombre de cycles de sélection et d'augmenter ainsi le nombre de séquences de fixation obtenues ([Roulet et al., 2002](#)), permettant de réaliser des modèles plus précis ([Nagaraj et al., 2008](#)).

Depuis la mise au point de la méthode SELEX, des avancées considérables ont été réalisées dans les techniques de séquençage, permettant l'obtention de millions de séquences à la fois : on parle de séquence haut-débit (*high-throughput*) ou encore séquençage massivement parallèle. L'utilisation de ces nouvelles techniques dans l'expérience SELEX a mené à la méthode

 1.4. *Les interactions protéine-ADN : mesures expérimentales*

HT-SELEX ([Nagaraj et al., 2008](#)), aussi appelée Bind-n-Seq ([Zykovich et al., 2009](#)). Il est alors possible d'estimer un modèle d'énergie à partir des fréquences d'observation des différentes séquences dès le premier cycle ([Nagaraj et al., 2008](#)). Des cycles supplémentaires permettent d'obtenir plus d'information sur les séquences les plus spécifiques, notamment sur la présence de contributions non indépendantes à l'énergie, ou de compenser la faible spécificité d'un TF. L'avantage de cette technique est que la taille des sites de fixation n'est pas limitée. Ainsi, avec une nanomole d'ADN ($\sim 10^{15}$ séquences) on peut couvrir l'ensemble des sites de 25bp possibles. Le séquençage haut-débit permet d'en échantillonner $\sim 10^8$, ce qui est largement suffisant pour contraindre des modèles d'énergie indépendants, même pour des TFs ayant des sites de fixations de taille $> 15\text{bp}$ comme c'est souvent le cas chez la bactérie. Cette technique a récemment été poussée encore plus loin ([Jolma et al., 2010](#)). En utilisant des protéines marquées, les auteurs ont réalisé un HT-SELEX à partir d'extraits cellulaires, et en ajoutant un code barre aux séquences d'ADN de chaque expérience, ils ont pu analyser les sites de fixation pour plusieurs TFs en parallèle. Ils ont ensuite utilisé cette technique pour obtenir des modèles de spécificité pour 411 TFs humains, la plus grande étude de ce genre réalisée à ce jour ([Jolma et al., 2013](#)).

1.4.2 Approche clonale : la technique de simple hybride

Contrairement aux approches précédentes, la technique de simple hybride (*Bacterial one-hybrid* ou B1H) n'est pas purement *in vitro*, au sens où l'interaction protéine-ADN est testée au sein d'une bactérie. Néanmoins, parce que l'interaction n'est pas testée dans son contexte cellulaire d'origine, nous la considérerons comme telle. Cette approche repose sur l'intégration par une bactérie hôte de deux vecteurs d'expression génétique, ou plasmides. Le premier exprime le facteur de transcription d'intérêt fusionné à une sous-unité de l'ARN polymérase (l'appât), c'est la protéine « hybride ». L'autre contient une région de séquence aléatoire représentant un site de fixation potentiel (la proie) en amont d'un promoteur à faible activité. La fixation de cette région par la protéine hybride permet l'activation d'un gène de sélection, généralement *HIS3*, un gène de la levure requis pour la biosynthèse de l'histidine et dont l'homologue bactérien est absent de la souche d'*Escherichia coli* utilisée. La croissance des cellules a lieu dans un milieu ne contenant pas l'histidine. Dans ces conditions, les bactéries n'exprimant pas *HIS3* ne peuvent croître. Ainsi, seules les bactéries au sein desquelles le facteur de transcription se fixe à la proie expriment *HIS3*, croissent et forment des colonies, d'où

Chapitre 1. Introduction générale.

la notion de gène de sélection. Par ailleurs, la stringence de la sélection peut être modulée en ajoutant au milieu différentes concentrations de 3-amino-triazole (3-AT), un inhibiteur de *HIS3*. De cette façon l'affinité du site de fixation peut être estimée plus finement.

Dans les études de ce type, les sites de fixation présents au sein des colonies sont séquencés individuellement, ce qui permet d'obtenir environ 50 séquences pour une expérience de sélection donnée. Néanmoins, il semble possible d'utiliser les nouvelles technologies de séquençage pour récupérer l'ensemble des sites de fixation des bactéries présentes sur une plaque ([Stormo and Zhao, 2010](#)). À l'instar de la méthode HT-SELEX, on obtient des millions de sites, ceux ayant une plus grande affinité étant présents à plusieurs centaines de milliers d'exemplaires, et ceux ayant une faible affinité étant présent en un seul voire aucun exemplaire.

Notons qu'il est aussi possible d'adopter la démarche inverse, c'est-à-dire de partir de quelques sites de fixation présumés fonctionnels mais pour lesquels on ne connaît pas le TF associé. En utilisant une bibliothèque de plasmides codant pour différents TFs hybrides, il est alors possible de déterminer si l'un d'entre eux possède une affinité importante avec les sites testés.

1.4.3 Approches *in vivo* : ChIP-on-chip, ChIP-seq, DNase I

Dans cette section, nous nous intéressons aux techniques permettant d'identifier les sites de fixation d'un facteur de transcription sur le génome. Ces méthodes se basent sur des extraits cellulaires (de 10^4 à 10^8 cellules) qui peuvent provenir d'un tissu homogène (un seul type de cellule) ou hétérogène (plusieurs types de cellules), voire de l'organisme entier si la dissection est impossible (embryon de mouche par exemple). L'information obtenue est donc toujours conditionnée par ce matériau de départ, et l'on n'obtient que les sites *accessibles* étant donnés le type cellulaire et la période de développement étudiés.

- **Immunoprécipitation de la chromatine : ChIP-on-chip et ChIP-seq**

La technique d'immunoprécipitation de la chromatine (ChIP) (fig. [1.11](#)) consiste dans un premier temps à induire la réticulation (*crosslink*) des protéines se liant à l'ADN en traitant les cellules avec de la formaldéhyde. Cette étape permet de transformer les liaisons faibles

1.4. Les interactions protéine-ADN : mesures expérimentales

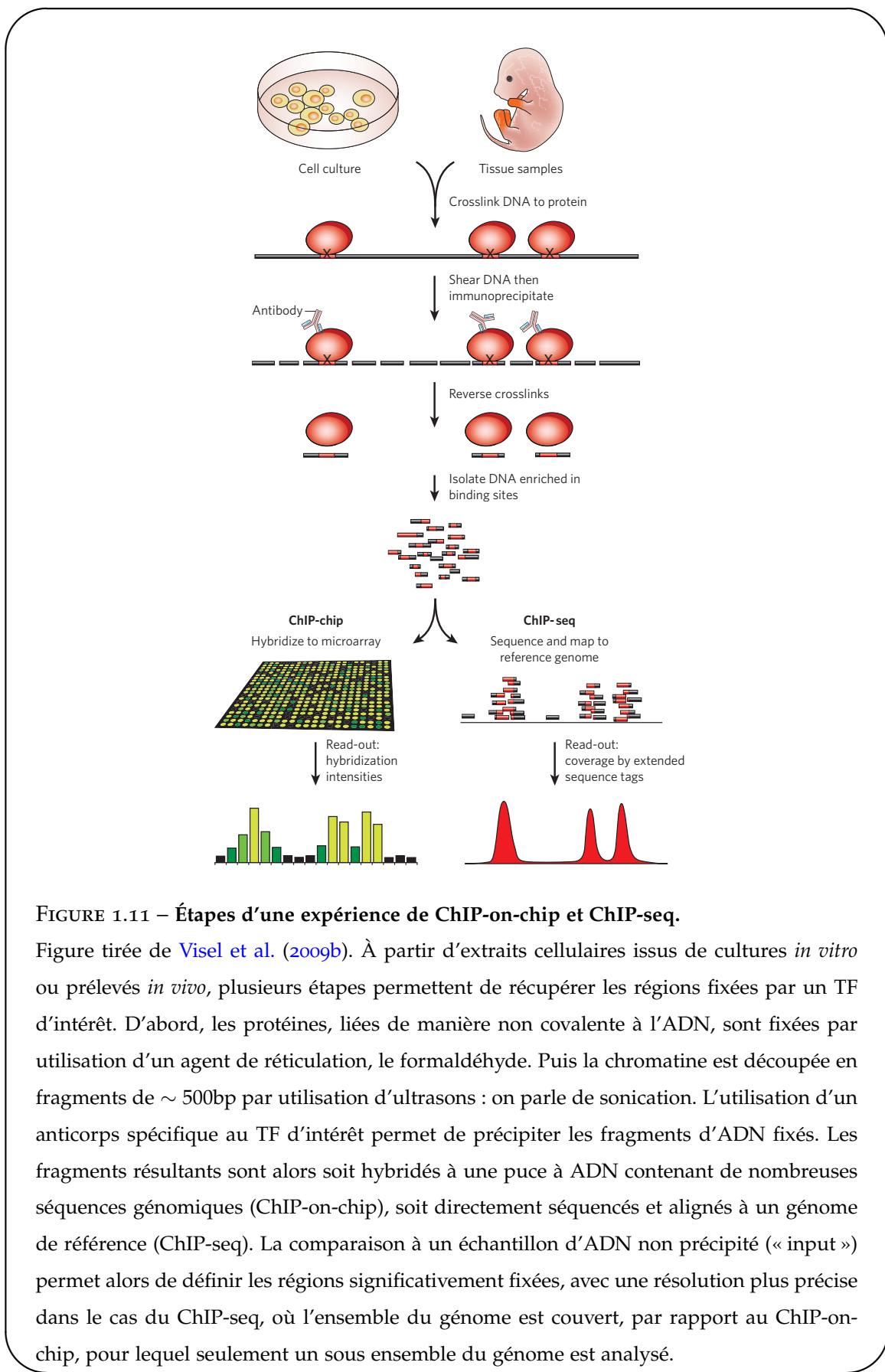


FIGURE 1.11 – Étapes d'une expérience de ChIP-on-chip et ChIP-seq.

Figure tirée de Visel et al. (2009b). À partir d'extraits cellulaires issus de cultures *in vitro* ou prélevés *in vivo*, plusieurs étapes permettent de récupérer les régions fixées par un TF d'intérêt. D'abord, les protéines, liées de manière non covalente à l'ADN, sont fixées par utilisation d'un agent de réticulation, le formaldéhyde. Puis la chromatine est découpée en fragments de ~ 500bp par utilisation d'ultrasons : on parle de sonication. L'utilisation d'un anticorps spécifique au TF d'intérêt permet de précipiter les fragments d'ADN fixés. Les fragments résultants sont alors soit hybrides à une puce à ADN contenant de nombreuses séquences génomiques (ChIP-on-chip), soit directement séquencés et alignés à un génome de référence (ChIP-seq). La comparaison à un échantillon d'ADN non précipité (« input ») permet alors de définir les régions significativement fixées, avec une résolution plus précise dans le cas du ChIP-seq, où l'ensemble du génome est couvert, par rapport au ChIP-on-chip, pour lequel seulement un sous ensemble du génome est analysé.

Chapitre 1. Introduction générale.

protéine-ADN en liaisons covalentes. Une fois les protéines fixées, la chromatine est découpée par digestion enzymatique ou en la soumettant à des ultrasons (c'est la sonication), résultant en des fragments de taille variant entre 200 et 600bp. Ces fragments sont ensuite immunoprécipités en présence d'un anticorps spécifique d'un facteur de transcription ou d'un isoforme d'histone (dans le cas d'une étude du paysage épigénétique) d'intérêt, permettant ainsi de récupérer tous les sites de fixation dans le génome. Après purification des fragments précipités, l'échantillon peut être analysé soit par hybridation sur puce (ChIP-on-chip) ou par séquençage haut débit (ChIP-seq).

Dans le cas du ChIP-on-chip, l'échantillon immunoprecipité et l'ADN de départ (*input*) sont marqués avec des colorants fluorescents et hybriderés sur une puce à ADN composée de très nombreux puits contenant des oligonucléotides (courtes séquences d'ADN) correspondant à différentes régions du génome. Dans le meilleur cas, ces oligonucléotides couvrent l'ensemble du génome. Les sites de liaison sont identifiés par l'écart d'intensité entre les signaux de fluorescence des conditions d'immunoprecipitation et d'*input*.

Dans le cas du ChIP-seq, l'échantillon immunoprecipité est analysé par séquençage à haut débit, résultant en une librairie de *reads* d'une longueur typique variant entre 27 et 50bp issus des extrémités des séquences. Ces *reads* sont ensuite alignés sur un génome de référence. À chaque position du génome correspond ainsi un certain nombre de séquences précipitées et d'*input*. En comparant ce nombre au nombre moyen dans le locus et à l'*input*, il est possible d'identifier des pics correspondant à la fixation du facteur (voir par exemple le programme d'appel de pics ChIP-seq MACS ([Zhang et al., 2008](#))).

Dans les deux cas, il faut noter que l'on a affaire à la fixation *moyenne* du facteur sur l'ADN dans la population de cellules étudiée. Ainsi, un petit pic peut représenter aussi bien une fixation forte dans un petit sous-ensemble de cellules (par exemple celles qui sont à un certain état d'avancement du cycle cellulaire) qu'une fixation moyenne dans l'ensemble de la population. L'expérience de ChIP-seq offre une résolution bien plus précise ($\leq 100\text{bp}$) que la méthode ChIP-on-chip (fig. 1.12). En effet, dans ce dernier cas la résolution est limitée par le nombre d'oligonucléotides utilisés, qui sont dans le meilleur des cas répartis sur le génome avec 35 – 100 nucléotides d'écart entre deux instances. Pour se comparer à la ChIP-seq, il

1.4. Les interactions protéine-ADN : mesures expérimentales

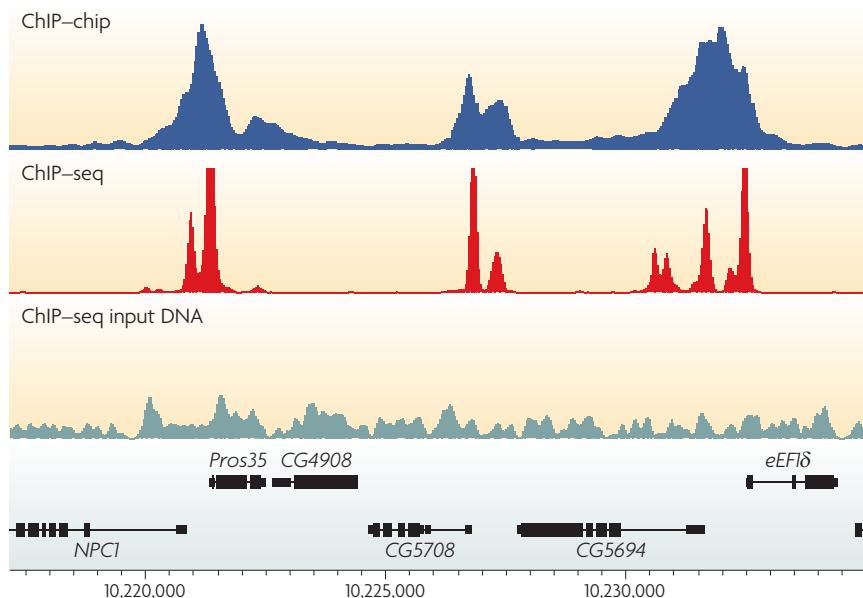


FIGURE 1.12 – Résolution des expériences ChIP-on-chip et ChIP-seq. Figure tirée de (Park, 2009), montrant les profils de fixation de la protéine Chromator générés à partir d'expériences de ChIP-on-chip (intensité relative par rapport au contrôle, bleu) et de ChIP-seq (densité de séquences, rouge) dans la lignée cellulaire S2 de *Drosophila melanogaster*. On peut noter la plus grande résolution de l'expérience ChIP-seq pour déterminer les sites de liaison. L'ADN utilisé en *input* de l'expérience de ChIP-seq et servant de contrôle est montré en gris, et les gènes du locus indiqués en noir.

faudrait que tous les oligonucléotides se superposent à une base près, ce qui demanderait un trop grand nombre de puces.

- **Empreinte à la DNase I (*DNase I footprinting*)**

Contrairement aux techniques précédentes, l'empreinte à la DNase I ne repose pas sur l'étude d'un facteur de transcription précis, mais permet au contraire d'obtenir l'ensemble des sites de fixation dans le génome pour un type cellulaire donné, avec une précision au nucléotide près. Cette méthode repose sur le fait que la fixation stable des facteurs de transcription à l'ADN n'est possible que si la région est pauvre en nucléosomes, les protéines autour desquelles s'enroule l'ADN : on parle de région de chromatine ouverte. Ces régions sont préférentiellement digérées par l'endonucléase DNase I. Étant donné que la majorité de l'ADN est enroulé autour de nucléosomes, les sites hypersensibles à la digestion par DNase I (*DNase I-hypersensitive* ou DHS) correspondent essentiellement à des régions de chromatine

Chapitre 1. Introduction générale.

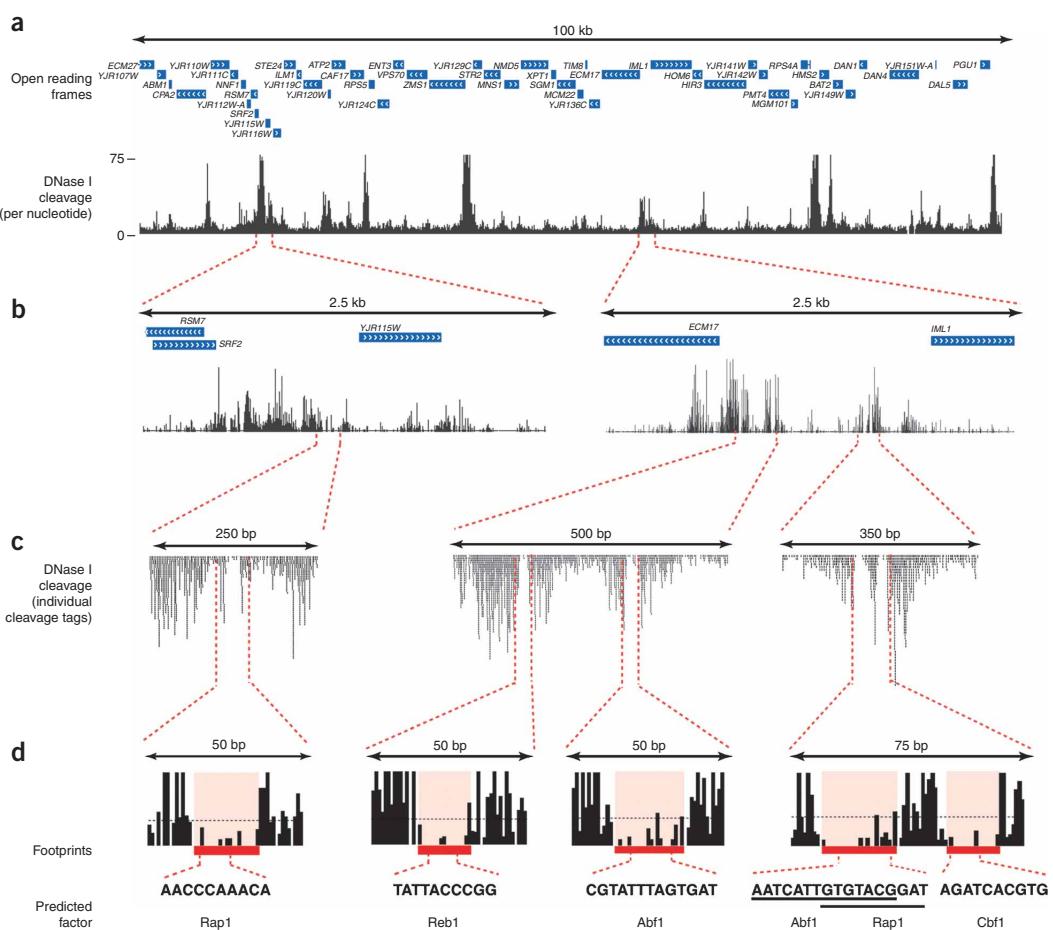


FIGURE 1.13 – Expérience d'empreinte à la DNase I chez la levure : vers une résolution au nucléotide près. Figure tirée de (Hesselberth et al., 2009). (a) Densité de digestion de la DNase I par nucléotide dans une région de 100kb du génome de la levure contenant ~ 50 gènes (boîtes bleues). On voit clairement certaines régions promotrices marquées par la DNase I. (b) Zoom sur deux régions de 2.5kb. (c) Zoom sur des régions de 250bp. Le nombre d'événements de digestion par nucléotide est marqué par des empilements de boîtes noires, révélant des régions protégées : ce sont les empreintes à la DNase I. (d) Les empreintes sont associées à des motifs de régulation de la levure connus. Les pointillés noirs indiquent le nombre moyen de digestion par nucléotide dans le génome (~ 2 digestions par base).

ouverte ayant des rôles de régulation génétique : promoteurs, enhancers...

En combinant la technique de DHS avec le séquençage à haut débit, l'expérience de DNase-seq permet d'identifier tous les types de région de régulation à l'échelle du génome (Thurman et al., 2012). Les régions riches en sites de digestion identifient alors les sites DHS. Par ailleurs, au sein d'un site DHS, il y a de petites régions ($\sim 15\text{bp}$) qui sont protégées de la digestion par DNase I : ce sont les empreintes à la DNase I ou *DNase I footprints* (fig. 1.13). Ces empreintes sont dues à la présence de protéines ou de complexes fixés à l'ADN. Cette technique de détection de sites de liaison par empreinte à la DNase I existe depuis 30 ans mais n'a que récemment été porté à l'échelle génomique. En comparant à des données ChIP-seq ou en utilisant des bases de données de motifs de facteurs de transcription, il est possible d'identifier le facteur correspondant dont les sites de fixation sont alors connus au nucléotide près.

1.5 Les modules de cis-régulation (CRMs)

Nous l'avons vu en section 1.2.2, les séquences d'ADN régulant l'expression génétique – CRMs pour *Cis-Regulatory Modules* – jouent un rôle prépondérant au cours du développement des organismes. Ces CRMs assurent en effet l'orchestration de l'expression de gènes spécifiques aux différentes étapes du développement et aux divers types cellulaires. Ils sont au cœur de l'évolution des réseaux génétiques, car ils dictent les interactions entre gènes. De plus, leur altération peut conduire à de nombreuses pathologies, liées pour la plupart à une expression génétique aberrante. Notamment, la majeure partie des variants génétiques qui sont associés de manière significative à une susceptibilité envers une maladie sont situés hors des régions codant pour des protéines, suggérant qu'un certain nombre affectent non pas la forme de la protéine engendrée mais l'expression du gène la produisant en détruisant une activité CRM. Dans cette partie, nous présentons les différents types de CRMs, leur structure, et leur évolution.

1.5.1 Les différents types de CRMs

Selon leur rôle dans la régulation de l'expression génétique, les CRMs peuvent être distingués en trois catégories.

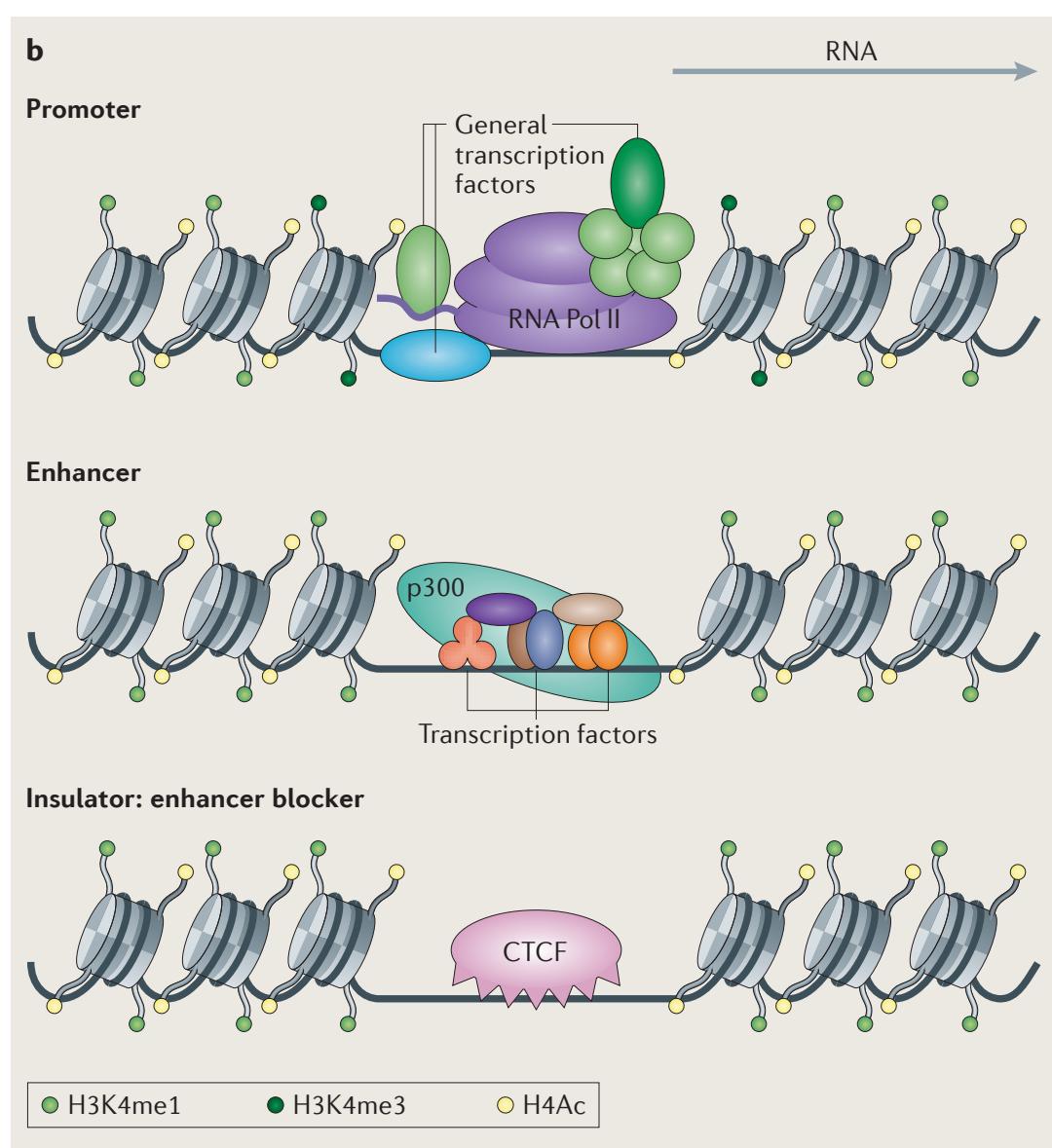


FIGURE 1.14 – Les différents types de CRMs et leurs marques épigénétiques.

Figure tirée de ([Hardison and Taylor, 2012](#)). La notion de CRM renvoie à un regroupement de sites de liaison pour un ou plusieurs facteurs de transcription. Les CRMs peuvent être regroupés en plusieurs classes : les promoteurs, les *enhancers/silencers*, et les insulateurs. Les CRMs des différentes classes partagent les marques d'acétylation H3Ac et H4Ac, les promoteurs actifs sont spécifiquement marqués par H3K4me3, et les enhancers et insulateurs par H3K4me1. Les enhancers sont par ailleurs souvent fixés par le co-activateur p300. Enfin, chez les mammifères les insulateurs recrutent CTCF pour bloquer l'activation par les enhancers.

- **Promoteurs**

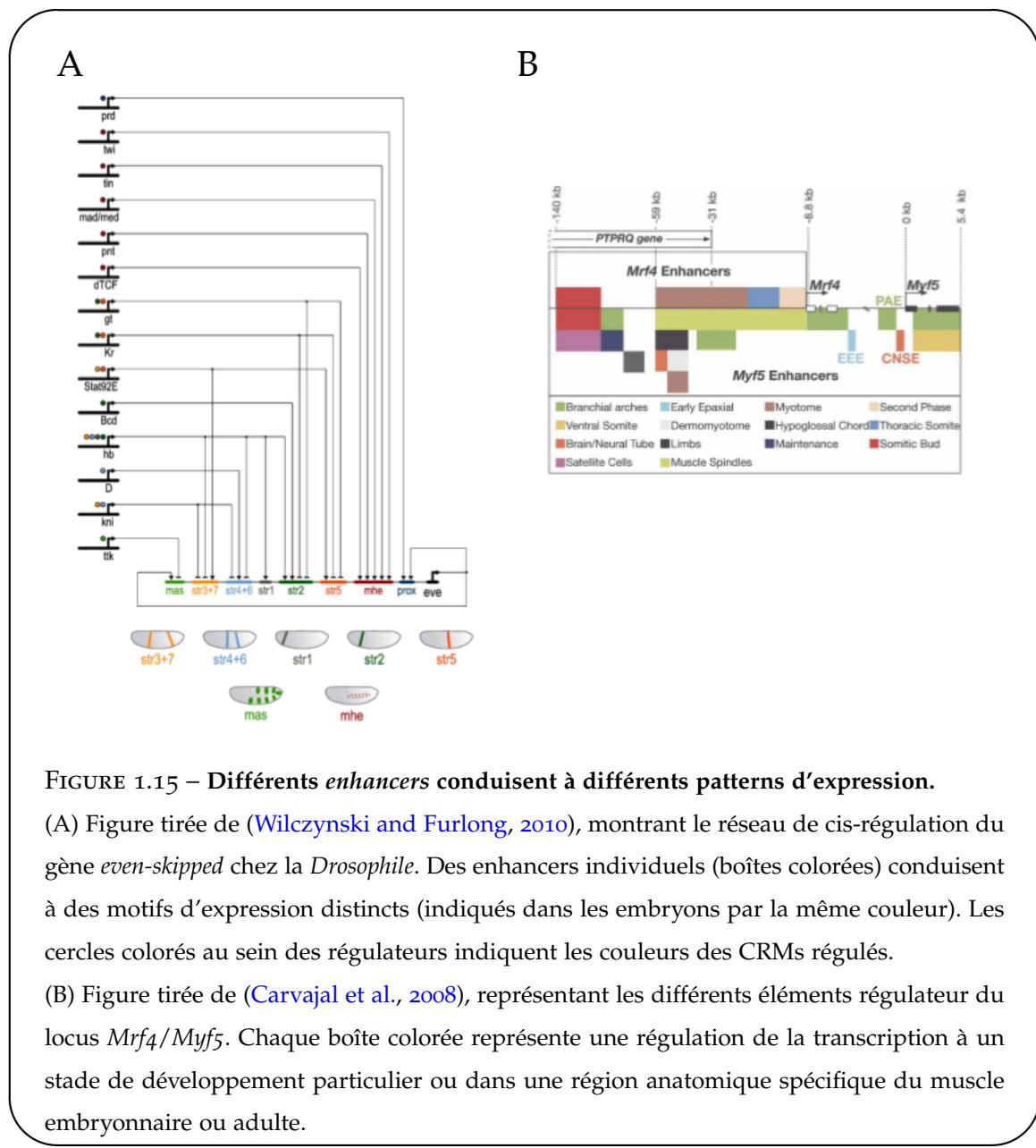
Les promoteurs permettent la fixation de l'ARN polymérase pour débuter la formation d'un transcript ARN au site d'initiation de transcription (*Transcription Start Site* ou TSS). Dans les promoteurs fixant l'ARN polymérase II (la majorité des promoteurs eucaryotes), des facteurs de transcription généraux se fixent à un cœur de $\sim 100\text{bp}$ autour du TSS afin de faciliter la fixation du complexe de polymérase. Ces coeurs de promoteurs contiennent pour certains des motifs stéréotypés, comme la boîte TATA, et ont un TSS bien déterminé ; néanmoins la plupart des promoteurs des génomes mammifères sont des régions riches en GC et en dinucléotides CpG (les « îlots CpG ») qui ne possèdent pas de boîte TATA et permettent l'initiation de la transcription dans un interval d'environ 100 bases (Carninci et al., 2006). Au niveau épigénétique, les promoteurs actifs sont caractérisés par une région pauvre en nucléosomes en amont du TSS, flanquée de nucléosomes possédant la marque de méthylation H3K4me3.

- ***Enhancers et silencers***

Les *enhancers* et *silencers* sont respectivement définis par leur effet positif ou négatif sur l'expression d'un gène cible. Cet effet peut notamment être observé par transfert d'un plasmide contenant l'élément régulateur en amont d'un gène rapporteur dans un animal transgénique ou dans des cultures cellulaires transfectées (voir 1.6.4). Leur activité ne dépend généralement pas de leur position et de leur orientation sur le plasmide. Selon l'environnement cellulaire, une région régulatrice peut être soit *enhancer* soit *silencer*, en fonction de la nature de co-activateurs ou de co-répresseurs des TFs recrutés. Il y a néanmoins relativement peu de *silencers* caractérisés et l'on utilise le terme d'*enhancers* pour désigner de manière générale ces régions régulatrices.

Les *enhancers* peuvent se situer à des distances variables du gène qu'ils régulent (Maniatis et al., 1987), pouvant parfois aller jusqu'à 1 Mb comme dans le cas de *Shh* chez la souris (Lettice et al., 2003) (voir fig. 1.24). Les enhancers contiennent de multiples sites de fixations de TFs. Cette multiplicité est requise pour l'activité enhancer, comme cela l'a été montré pour le premier enhancer découvert : celui du virus simien 40 (SV40) (Schirm et al., 1987; Ondek et al., 1988). Un gène peut par ailleurs posséder plusieurs enhancers distincts conduisant à des expressions spécifiques dans différents tissus, comme cela l'a été montré dans le cas du gène *eve* chez la *Drosophila* (Wilson and Odom, 2009) ou dans le cas du cluster de gènes de détermination myogénique *Myf5* et *Mrf4* chez les mammifères (Carvajal et al., 2008) (fig. 1.15). Ainsi,

Chapitre 1. Introduction générale.



les différents enhancers d'un même gène peuvent être vus comme autant de points d'entrée d'un réseau de régulation génétique, représentant diverses fonctions logiques et intégrant différentes information spatio-temporelles pour produire en sortie une expression génétique spatio-temporelle finement contrôlée (Bolouri and Davidson, 2002; Buchler et al., 2003).

Enfin, comme décrit en fig. 1.14, les enhancers sont associés à de hauts niveaux de marque épigénétique H3K4me1 (Heintzman et al., 2009) et sont souvent fixés par le co-activateur p300 (Wang et al., 2005; Heintzman et al., 2009).

- **Insulateurs**

Les insulateurs sont des CRMs qui restreignent l'effet des enhancers sur leur gène cible ([Wallace and Felsenfeld, 2007](#)). Ainsi, certains insulateurs possèdent une activité de blocage d'enhancers. Situés entre un enhancer et un promoteur cible, ces insulateurs bloquent l'activité de l'enhanсer, conduisant à une réduction de l'expression du gène cible ([Chung et al., 1993](#)). Chez les mammifères, la fixation de la protéine CTCF est nécessaire à cette activité de blocage de l'activité enhancer ([Bell et al., 1999](#)), alors que chez la *Drosophila* et plusieurs autres insectes il existe au moins quatre protéines additionnelles qui sont suffisantes à la réalisation de cette activité ([Schoborg and Labrador, 2010](#)). Par ailleurs, les insulateurs peuvent servir de barrière de protection contre des marques d'hétérochromatine répressives. De tels insulateurs permettent notamment d'éviter les effets de positions – la modification de l'expression d'un gène selon sa position dans le chromosome – lorsqu'ils entourent un gène rapporteur intégré au hasard dans le génome ([Recillas-Targa et al., 2002](#)). Cette activité passe notamment par le recrutement de *USF*, protéine qui recrute des enzymes de modification de la chromatine. Un insulateur peut combiner les activités de barrière de protection et de blocage d'enhanсer.

De même que les enhancers, les insulateurs peuvent se situer à des distances variables des gènes qu'ils régulent. Il est à noter que la protéine CTCF possède d'autres fonctions que celle d'isolation, et tous les sites de CTCF ne correspondent pas forcément à des insulateurs ([Philips and Corces, 2009](#)).

1.5.2 Grammaire des enhancers : enhanceosome vs billboard

Nous l'avons vu, les CRMs contiennent en général de multiples sites de liaisons (TFBS) pour un ou plusieurs TFs. On parle de *clustering* (regroupement). Lorsque les TFBS correspondent à plusieurs TFs différents, on parle de CRM hétérotypique, et dans le cas où ils correspondent à un même TF, on parle de CRM homotypique. Cette distinction est surtout utile pour décrire les différentes méthodes de prédiction de CRM, car la plupart des CRMs identifiés chez les Métazoaires sont hétérotypiques ([Aerts, 2012](#)). L'organisation de ces sites de liaison relève de deux types d'architecture principaux (fig. [1.16](#)).

- **Le modèle “enhanceosome”**

Dans ce modèle, l'architecture des sites de liaison est de prime importance. Le paradigme en est l'enhanсer du gène humain interferon- β , sur lequel 8 TFs se fient pour former une

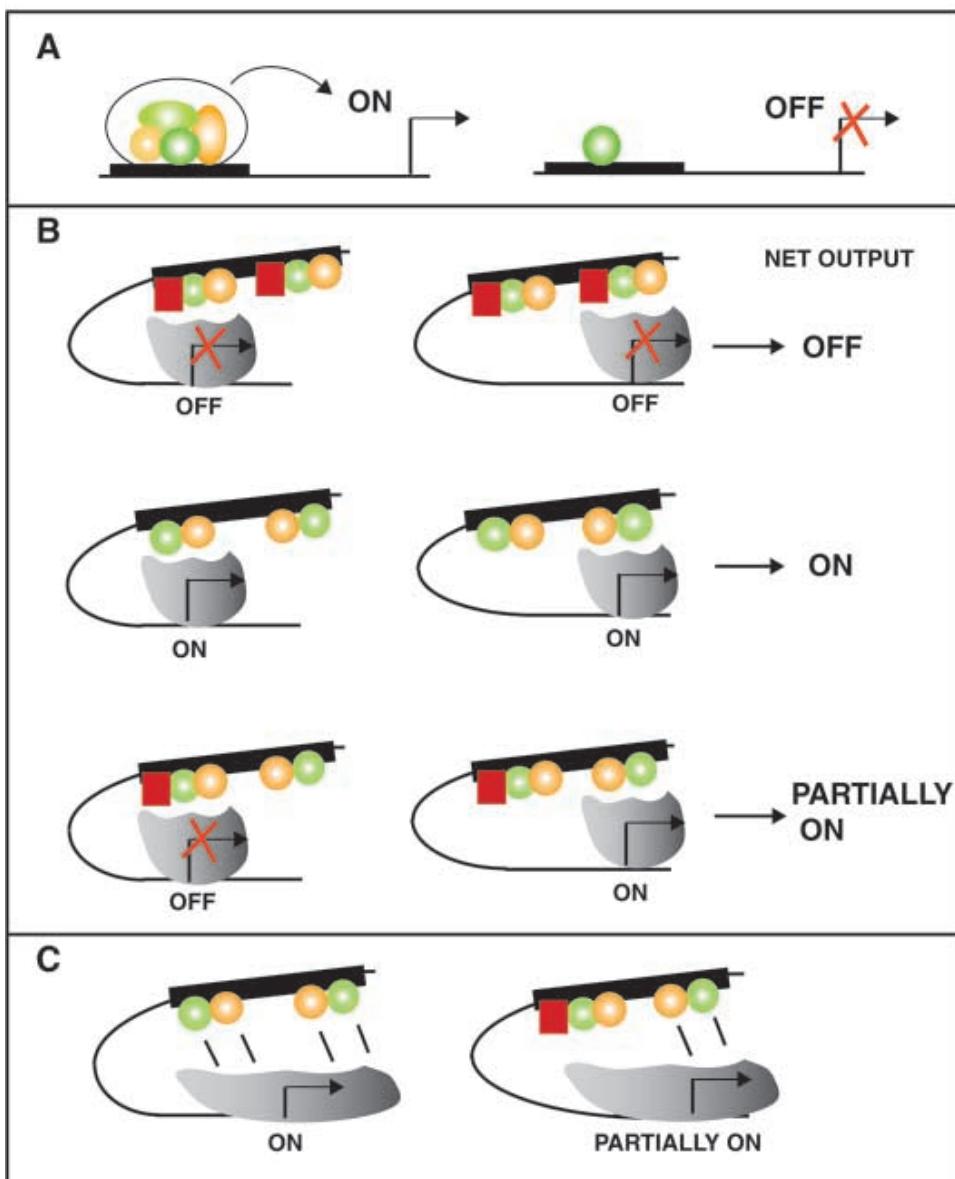


FIGURE 1.16 – Deux modèles d’enhancers : enhanceosome et billboard.

Figure tirée de ([Kulkarni and Arnosti, 2003](#)). (A) Dans le modèle enhanceosome, l’enhancer traite l’information des multiples TFs qui le fixent. Un complexe très structuré crée une interface qui recrute la machinerie de transcription basale. L’enhancer peut être vu comme un ordinateur moléculaire qui produit à partir d’entrées multiples un seul signal vers la machinerie de transcription. Le gène cible n’est activé qu’en cas de formation du complexe entier, ce qui fournit un interrupteur binaire on/off seulement activé en cas de stimulus adéquat. La déstabilisation du complexe en changeant par exemple la concentration d’une des protéines permettrait alors d’obtenir une réponse graduelle. (B,C) Modèle d’enhancer « billboard ». Dans ce cas, l’enhancer ne consiste pas en une seule unité de régulation, mais en des sous-unités pouvant contenir différentes informations (répression ou activation par exemple) que la machinerie basale échantillonne soit itérativement (B), soit simultanément (C).

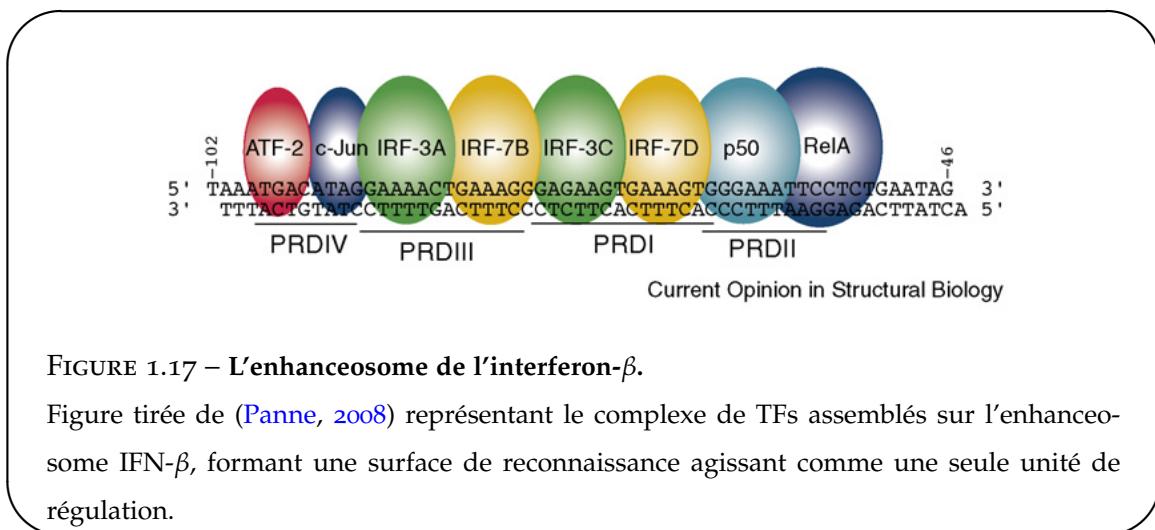


FIGURE 1.17 – L’enhanceosome de l’interferon- β .

Figure tirée de ([Panne, 2008](#)) représentant le complexe de TFs assemblés sur l’enhanceosome IFN- β , formant une surface de reconnaissance agissant comme une seule unité de régulation.

surface de reconnaissance continue ([Panne, 2008](#)). Les TFBS de cet enhancer se recouvrent les uns les autres, créant au final un complexe de TFs fixés à l’ADN agissant comme une seule unité de régulation (fig. 1.17).

- **Le modèle “billboard”**

La majorité des CRMs adhèrent à ce type d’organisation. L’architecture y est libre : les sites de liaisons n’ont pas de contrainte de nombre, d’ordre, de sens, ou d’espacement ([Kulkarni and Arnosti, 2003](#)). De tels CRMs sont propices à une détection informatique basée sur la densité en sites de liaisons pour différents TFs.

1.5.3 Évolution des enhancers

La fonction centrale que jouent les enhancers dans la régulation de l’expression génétique laisse à penser que ceux-ci seraient sous sélection et leur séquence serait donc plus conservée que celle des régions non codantes du génome. De fait, la comparaison de séquences non-codantes entre espèces proches s’avère être un mode de détection puissant des régions de régulation ([Prabhakar et al., 2006](#)). Ainsi, l’utilisation de la conservation entre des espèces lointaines comme l’homme et le poisson *Fugu* ou de l’extrême conservation entre des espèces proches comme l’homme, la souris et le rat, permet de détecter des régions ayant une activité enhancer *in vivo* avec un succès proche de 50% ([Pennacchio et al., 2006](#)). À l’instar de la régulation de l’interféron- β , de telles séquences très contraintes obéissent à une logique de type « enhanceosome » où la fonction est intimement liée à la séquence.

Contrastant avec cette vision d’enhancers très contraints, plusieurs études pointent vers

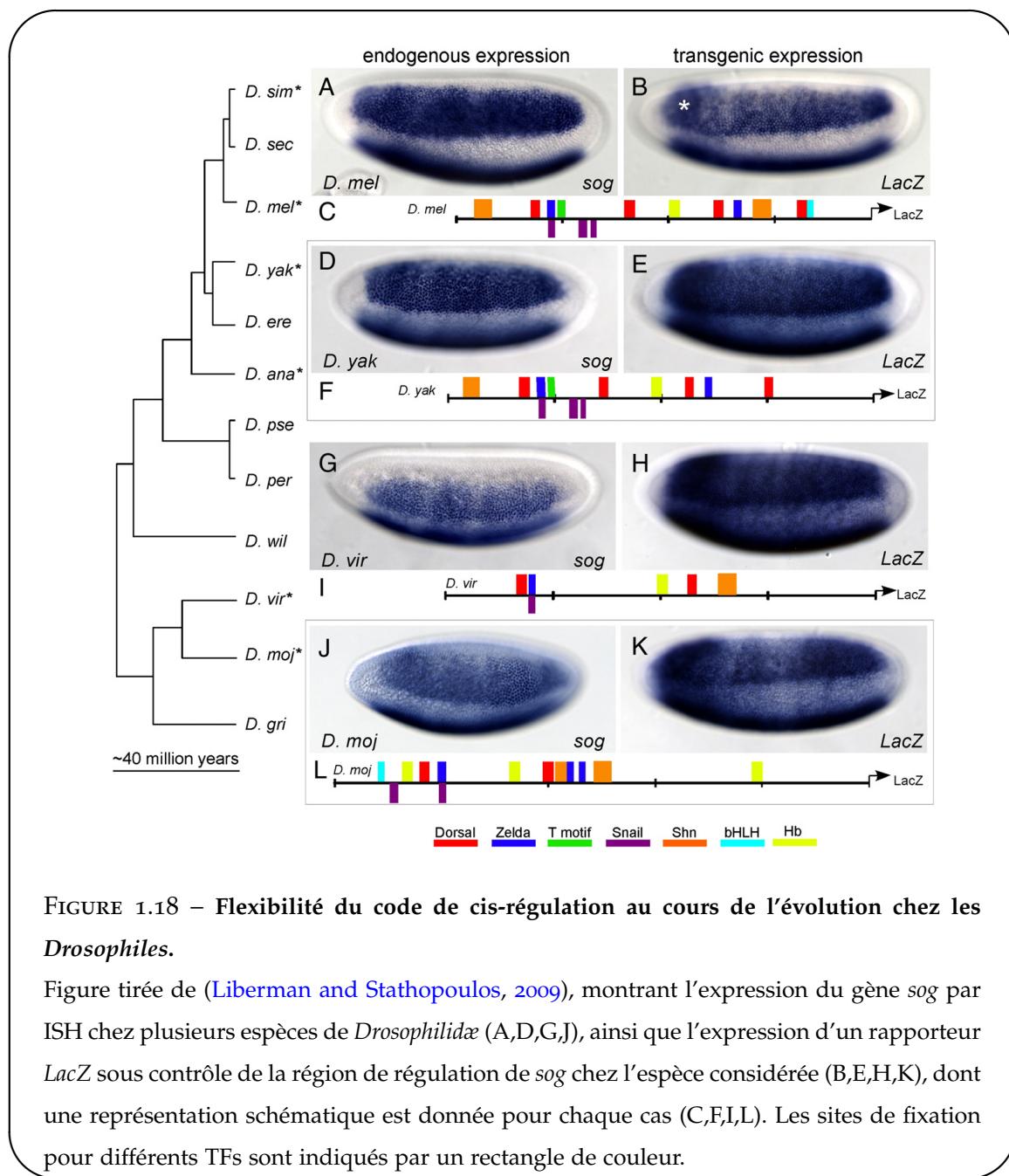


FIGURE 1.18 – Flexibilité du code de cis-régulation au cours de l'évolution chez les *Drosophiles*.

Figure tirée de ([Liberman and Stathopoulos, 2009](#)), montrant l'expression du gène *sog* par ISH chez plusieurs espèces de *Drosophilidae* (A,D,G,J), ainsi que l'expression d'un rapporteur *LacZ* sous contrôle de la région de régulation de *sog* chez l'espèce considérée (B,E,H,K), dont une représentation schématique est donnée pour chaque cas (C,F,I,L). Les sites de fixation pour différents TFs sont indiqués par un rectangle de couleur.

une plus grande flexibilité des séquences enhancers ([Ludwig et al., 2000](#); [Dermitzakis and Clark, 2002](#); [Moses et al., 2006](#)). Supportant l'idée que la plupart des enhancers se comportent selon le modèle « billboard », la grammaire des sites de fixation dans des séquences orthologues apparaît comme étant loin d'être figée ([Liberman and Stathopoulos, 2009](#)). Ainsi, l'enhancer régulant le gène *short gastrulation* (*sog*), bien que présentant chez différentes espèces de *Drosophiles* une architecture variable des sites de fixation le composant, conduit à un même motif d'expression (fig. 1.18). Cette idée qu'une panoplie de grammaires conduisent à une même régulation est confortée par les résultats de [Zinzen et al. \(2009\)](#) où des enhancers ayant des « entrées » différentes (i.e étant fixés par des TFs différents pendant des durées variables) produisent des « sorties » similaires, dans ce cas une expression spécifique à un tissu donné.

Supportant l'idée d'une flexibilité de la régulation, plusieurs études ont exhibé l'évolution rapide des sites de liaison de TFs dans le génome ([Wilson and Odom, 2009](#)). Une étude de la fixation génomique des facteurs de transcription CEBP α et HNF4 α dans les cellules du foie de 5 espèces de vertébrés (l'homme, deux espèces de souris, le chien et le poulet) a notamment montré que les événements de fixation conservés chez les 5 espèces sont très rares ($\sim 0.3\%$ des pics humains) et correspondent à des régions ultraconservées proches de gènes importants dans la spécification du foie ([Schmidt et al., 2010](#)). Par ailleurs, lors de la perte de fixation dans l'une des espèce, un gain de fixation proche ($\pm 10\text{kb}$) est observé dans la moitié des cas. Étonnamment, ces changements rapides du câblage du réseau affectent peu l'expression génétique globale ([Tirosh et al., 2008](#); [Odom et al., 2007](#)).

Cette évolution est en grande partie due à une évolution de séquence de fixation. Ainsi, une étude récente a utilisé une souris portant le chromosome 21 de l'homme pour comparer la fixation du facteur HNF4 α dans un contexte murin par rapport au contexte original ([Wilson et al., 2008](#)). Le paysage de fixation sur le chromosome 21 exogène a très précisément récapitulé celui observé chez l'homme (fig. 1.19), montrant que le contexte cellulaire est sensiblement le même chez les deux espèces. Par ailleurs, des modifications épigénétiques ainsi que l'expression des ARNm ont pu être récapitulées.

Reste la question du mécanisme permettant cette évolution rapide. Une étude portant sur 7 facteurs de transcription chez les mammifères a montré qu'une proportion importante ($\sim 20\%$) des régions de fixation de ces TFs se situent au sein de différentes familles de transposons ([Bourque et al., 2008](#)) (fig. 1.19). Ces transposons, ou éléments transposables, sont des anciens rétrovirus intégrés dans les génomes mammifères ayant la capacité de se dupliquer pour

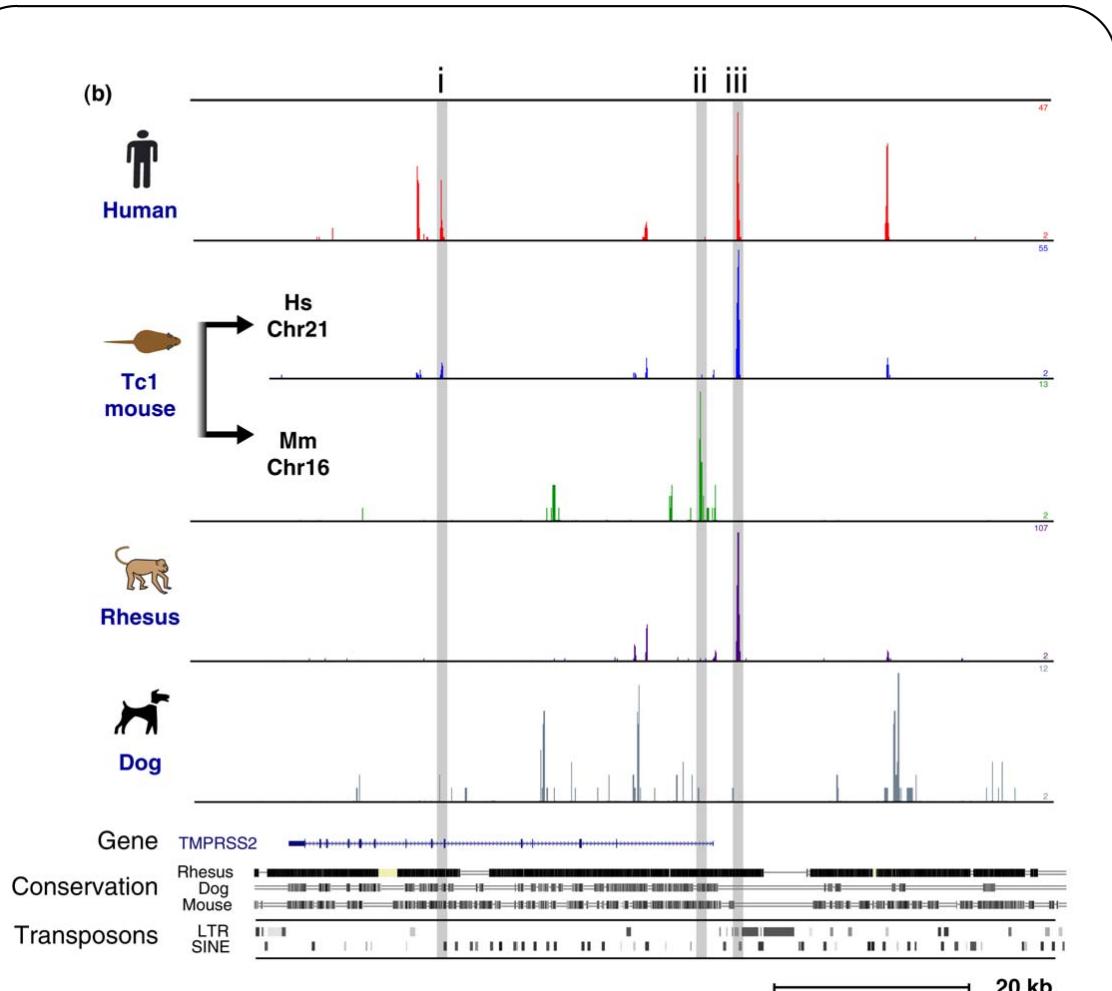


FIGURE 1.19 – Évolution de la fixation de HNF4α chez les mammifères.

Figure tirée de (Wilson and Odom, 2009), représentant la fixation par ChIP-seq (pics de couleur) du facteur de transcription humain HNF4α (ou son homologue murin HNF4a) chez l'homme, la souris, le macaque, le chien, ainsi qu'une souris transgénique contenant le chromosome 21 humain. Les zones grisées indiquent : (i) une fixation de HNF4α chez l'homme retrouvée sur le chromosome 21 humain de la souris mais pas chez le macaque malgré la proximité de séquence, (ii) une fixation de HNF4a spécifique à la souris, et (iii) une fixation spécifique aux primates qui a lieu sur des éléments transposables.

s'intégrer dans une autre région du génome et jouent un rôle fondamental dans l'évolution des génomes ([Cordaux and Batzer, 2009](#)). Leur accumulation dans le génome a vraisemblablement permis d'obtenir un matériau de base permettant de produire par mutations ponctuelles des éléments de régulation *de novo* ([Feschotte, 2008](#)). Par ailleurs, les transposons peuvent permettre de diffuser par « copier-coller » des éléments de régulation existant. Ainsi, des vagues d'expansion de transposons spécifiques à différentes espèces de mammifères sont à l'origine de la variabilité des régions de fixation observée dans le cas du facteur CTCF ([Schmidt et al., 2012](#)).

1.5.4 Les « shadow enhancers »

L'évolution des éléments de cis-régulation est un mécanisme majeur permettant la diversité animale. Néanmoins, de tels changements pourraient compromettre certaines activités génétiques essentielles. Des expériences de ChIP-on-chip ont suggéré que plusieurs gènes de développement actifs lors du développement précoce de l'embryon de Drosophile possèdent des CRMs secondaires, qui conduisent à des motifs d'expression génétique comparables à ceux produits par des CRMs « primaires » plus proximaux ([Zeitlinger et al., 2007](#)). L'expression de « shadow enhancer » a été proposée par Michael Levine en 2008 pour décrire ces CRMs redondants et souvent distaux de plusieurs dizaines de kb du gène régulé ([Hong et al., 2008](#)). Il est probable que de tels CRMs soient apparus au cours de l'évolution par duplication du CRM primaire, à l'instar du phénomène de duplication des séquences codant pour des protéines. L'avantage évident que peut conférer la redondance d'un élément de régulation est d'offrir de la robustesse face aux mutations. Par ailleurs, une telle redondance permet de faciliter la divergence et donc la spécialisation des différents CRMs. Ainsi les « shadow enhancers » semblent évoluer plus rapidement que les CRMs primaires auxquels ils sont apparentés ([Hong et al., 2008](#)) pour fournir de nouveaux sites de fixation et conduire à de nouvelles activités de régulation sans bloquer la fonction critique de certains gènes de développement.

Un exemple mêlant robustesse et divergence est le cas des multiples CRMs régulant le gène *Svb* chez la Drosophile. Chaque CRM est lié à la production d'un motif distinct de trichomes (excroissances de l'épithélium comparables à des poils) sur la larve : ainsi, plusieurs mutations dans ces différents CRMs sont nécessaires pour observer un changement morphologique conséquent ([McGregor et al., 2007](#)). Dans ce même système, il a été montré que deux CRMs supplémentaires, des « shadow enhancers », sont dispensables dans des conditions de

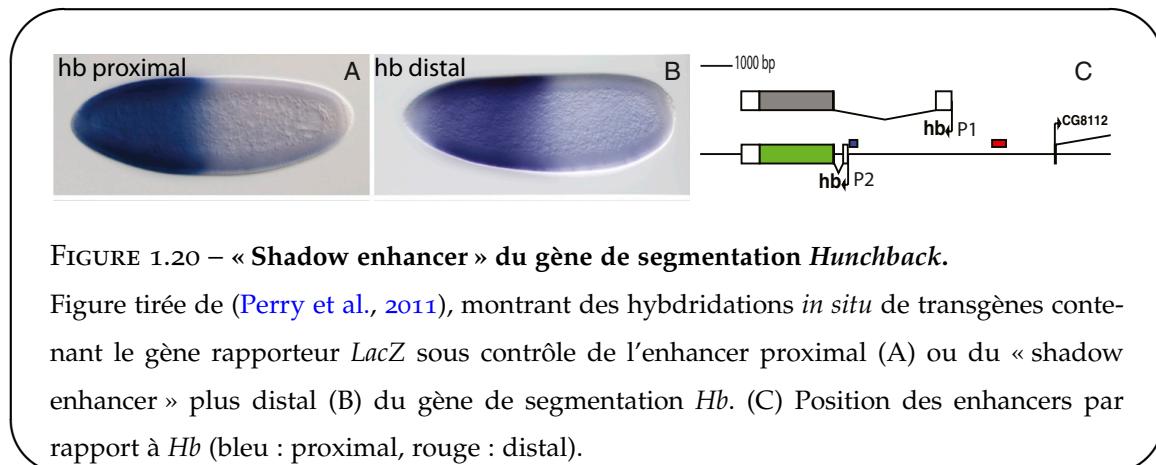


FIGURE 1.20 – « Shadow enhancer » du gène de segmentation *Hunchback*.

Figure tirée de ([Perry et al., 2011](#)), montrant des hybrides in situ de transgènes contenant le gène rapporteur *LacZ* sous contrôle de l'enhancer proximal (A) ou du « shadow enhancer » plus distal (B) du gène de segmentation *Hb*. (C) Position des enhancers par rapport à *Hb* (bleu : proximal, rouge : distal).

température usuelles, mais requis lorsque les embryons se développent dans des conditions de température extrêmes ([Frankel et al., 2010](#)).

Par ailleurs, il a été montré que les gènes de segmentation (ou gènes *gap*) de la Drosophile possèdent tous des « shadow enhancers » (fig. 1.20). Leur rôle semble être d'assurer une plus grande précision spatiale du motif d'expression du gène régulé : la perte de l'un des CRMs, proximal aussi bien que « shadow », conduisant à une expression trop restreinte ou trop répandue spatialement selon le cas ([Perry et al., 2011](#)).

1.5.5 Par delà les enhancers : les « super-enhancers »

Récemment, il a été montré que certains groupements d'enhancers peuvent agir comme une même unité de régulation : on parle de *super-enhancers* ([Whyte et al., 2013](#)). Ces régions de taille typique $\sim 10\text{kb}$ (fig. 1.21), sont fixées par des TFs maîtres et sont associées à des gènes encodant des régulateurs clés de l'identité cellulaire. Identifiés dans les cellules souches embryonnaires (ESCs), ces ensembles d'enhancers sont fixés par le complexe co-activateur Mediator, qui interagit avec la cohésine pour former un anneau permettant de connecter la région de régulation au promoteur ([Kagey et al., 2010](#)). Par ailleurs, les gènes associés aux super-enhancers possèdent un niveau particulièrement élevé d'expression et leur knock-down est associé à une perte de l'état souche des cellules.

Ainsi, ce second niveau d'organisation de la régulation pourrait simplifier la modélisation de la régulation du type cellulaire, en passant de millier de traces de fixation pour différents TFs à quelques centaines de super-enhancers contrôlant les gènes clés de l'identité cellulaire.

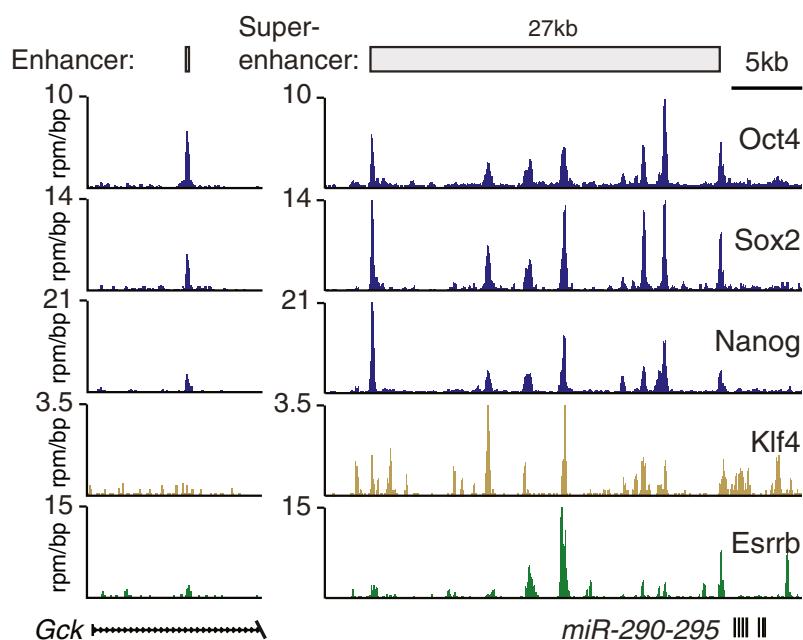


FIGURE 1.21 – De l'enhancer au super-enhancer.

Figure tirée de ([Whyte et al., 2013](#)), montrant les profils de ChIP-seq des TFs maîtres Oct4, Sox2, Nanog, Klf4 et Esrrb aux loci de *Gck* et *miR-290-295* dans les cellules souches embryonnaires. Le super-enhancer se distingue du simple enhancer par sa taille (27kb), sa grande concentration en TFs maîtres, notamment Klf4 et Esrrb, et la fixation de la protéine Med1 du complexe Mediator.

1.6 Prédiction et validation des CRMs

1.6.1 Méthodes utilisant la concentration en sites de fixation

Nous l'avons vu, une propriété des CRMs est leur grande concentration en TFBS. Ceci a motivé des approches de prédiction de promoteurs et d'enhancers basées sur leur contenu ou *clustering* en motif (fig. 1.22a). L'avantage de telles approches est qu'elles peuvent être réalisées avec seulement la séquence d'ADN génomique et des modèles de TFs ou motifs (par exemple des PWMs, voir fig. 1.10) représentant les facteurs de transcription impliqués dans le processus étudié. Cependant, les clusters de motifs sont très répandus dans les grands génomes, et sans l'ajout d'informations supplémentaires comme les marques épigénétiques ou l'expression des gènes voisins, ces approches produisent un grand nombre de faux positifs (éléments prédis comme positifs mais étant en réalité négatifs). Par ailleurs, les TFs impliqués ne sont pas toujours connus, et il faut alors apprendre des motifs putatifs à partir de séquences fonctionnelles.

- **Approches utilisant des motifs connus**

L'une des premières investigations basée sur le regroupement de TFBS utilisait 5 motifs connus de la détermination musculaire pour prédire par régression linéaire les CRMs actifs dans le muscle ([Wasserman and Fickett, 1998](#)). Le taux de validation était relativement bas, autour de 20%. De même, chez *Drosophila melanogaster*, plusieurs études ont utilisé le clustering de motifs pour prédire des CRMs de différents processus développementaux (par ex [Berman et al. \(2002\)](#)). Ces études ont trouvé de nouveaux enhancers validés expérimentalement (bonne sensibilité) mais avaient des taux de prédiction relativement bas, entre 15 et 30%. L'algorithme *Ahab* ([Rajewsky et al., 2002](#)), utilisant un modèle thermodynamique de fixation des TFs sur les CRMs, a quant à lui réussi à prédire un nombre bien plus important de régions fonctionnelles : ~ 80% des modules prédis à proximité de 29 gènes de segmentation chez la drosophile ont effectivement récapitulé le motif d'expression du gène associé ([Schroeder et al., 2004](#)). Ce succès semble notamment être dû au fait que ce modèle thermodynamique, basé sur une prise en compte exhaustive de toutes les segmentations possibles des CRMs en motifs et en ADN « background », permet de donner plus de poids au cas où plusieurs sites de faibles affinité pour un TF se trouvent au sein d'un même module, alors que les autres méthodes utilisent généralement un seuil de probabilité relativement élevé (afin d'éviter les faux posi-

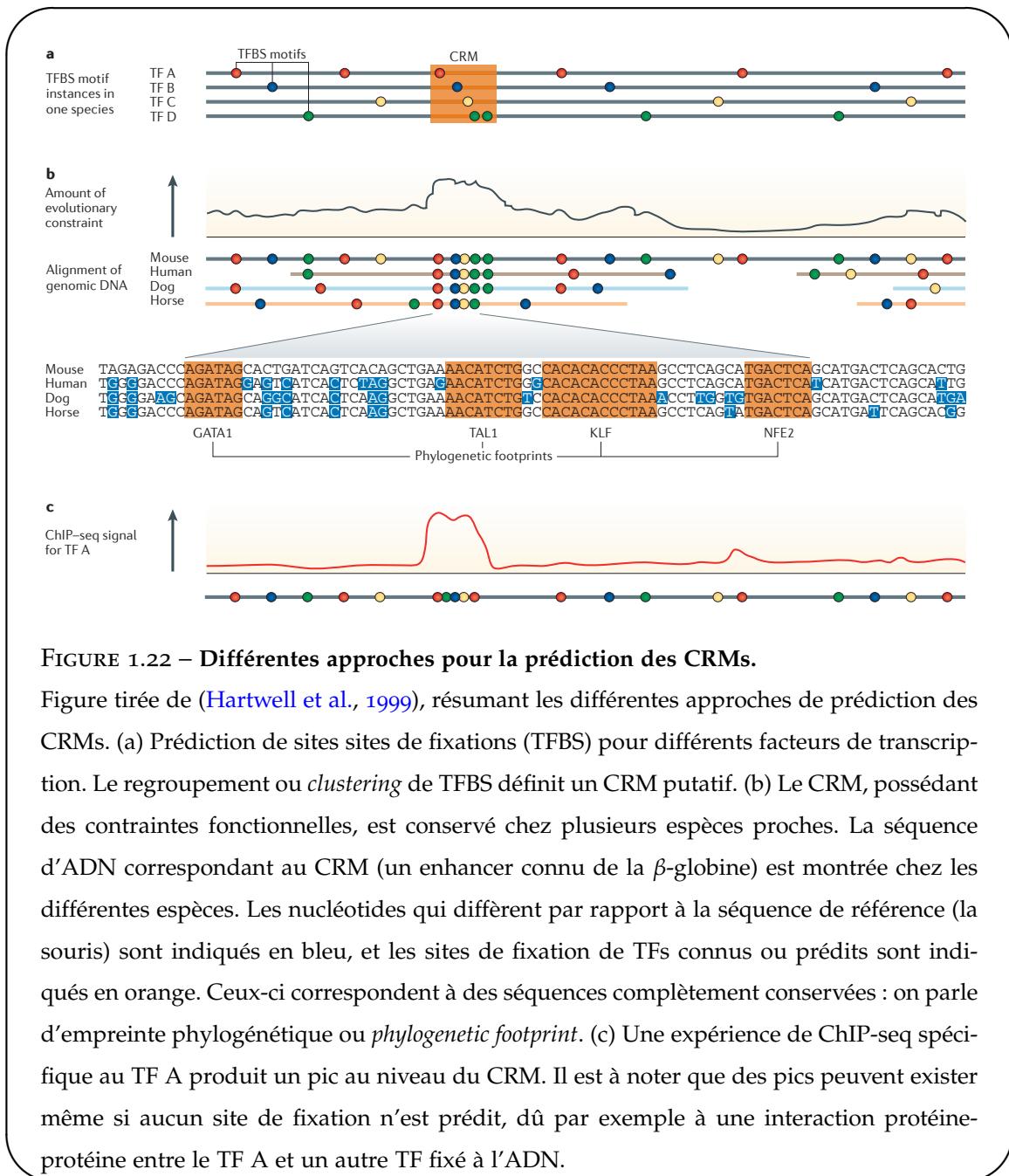


FIGURE 1.22 – Différentes approches pour la prédition des CRMs.

Figure tirée de (Hartwell et al., 1999), résumant les différentes approches de prédition des CRMs. (a) Prédiction de sites sites de fixations (TFBS) pour différents facteurs de transcription. Le regroupement ou *clustering* de TFBS définit un CRM putatif. (b) Le CRM, possédant des contraintes fonctionnelles, est conservé chez plusieurs espèces proches. La séquence d'ADN correspondant au CRM (un enhancer connu de la β -globine) est montrée chez les différentes espèces. Les nucléotides qui diffèrent par rapport à la séquence de référence (la souris) sont indiqués en bleu, et les sites de fixation de TFs connus ou prédits sont indiqués en orange. Ceux-ci correspondent à des séquences complètement conservées : on parle d'empreinte phylogénétique ou *phylogenetic footprint*. (c) Une expérience de ChIP-seq spécifique au TF A produit un pic au niveau du CRM. Il est à noter que des pics peuvent exister même si aucun site de fixation n'est prédit, dû par exemple à une interaction protéine-protéine entre le TF A et un autre TF fixé à l'ADN.

Chapitre 1. Introduction générale.

tifs) à partir duquel une séquence est considérée comme fixée par un TF. Par ailleurs, cette étude s'est restreinte à un ensemble de gènes connus pour lesquels les régions à proximité riches en TFBS ont *a priori* plus de chances d'être fonctionnelles. De manière générale, plus le domaine de recherche est étendu (par exemple, le génome entier), plus le nombre de faux positifs augmente.

- **Approches *de novo* où les motifs ne sont pas connus**

Lorsque les motifs (PWMs) ne sont pas connus à l'avance, il faut les générer *de novo* à partir de leur surreprésentation dans des CRMs connus. Par exemple, l'algorithme CisModule permet de générer des motifs et des modules simultanément en utilisant un modèle de mélange hiérarchique ([Zhou and Wong, 2004](#)). Lorsqu'il est appliqué aux CRMs musculaires introduits précédemment, il permet de retrouver certains motifs connus et permet de retrouver $\sim 70 - 80\%$ des séquences connues lorsqu'elles sont mélangées avec un nombre similaire de séquences aléatoires. Par ailleurs, l'apprentissage de modèles permettant de discriminer différentes classes de CRMs entre elles plutôt qu'une classe de CRMs par rapport à des séquences aléatoires ou intergéniques peut s'avérer plus fructueux. Ainsi, ([Smith et al., 2006](#)) ont utilisé des motifs connus ainsi que des motifs appris *de novo* avec le programme DME ([Smith et al., 2005](#)) pour leur capacité à discriminer des séquences appartenant à différents jeux de données de régions promotrices pour bâtir un modèle de régression logistique permettant de prédire l'activité tissu-spécifique dans 45 des 56 tissus humains et murins considérés. Il existe aussi plusieurs méthodes qui n'utilisent pas de motifs du type PWM, mais de purs modèles probabilistes tels que des chaînes de Markov d'ordre 5 ou des regroupements de « mots » de k nucléotides ou k -mers selon des critères de distance de Hamming et surreprésentés dans les séquences d'intérêt, par exemple ([Cao et al., 2010a](#)). Ces méthodes sont passées en revue dans ([Kantorovitz et al., 2009](#)), et elles peuvent atteindre des sensibilités de $\sim 60\%$ pour la prédiction de CRMs mammifères. L'intérêt est que ces études ne présument pas d'un modèle de fixation des TFs à l'ADN. C'est aussi un désavantage, puisqu'elles sont moins informatives quant au réseau génétique sous-jacent et aux mécanismes de régulation impliqués.

1.6.2 Méthodes utilisant la phylogénie

Les approches utilisant la comparaison des génomes de différentes espèces pour prédire des CRMs sont basées sur l'idée que les séquences de régulation sont plus fortement conser-

vées que l'ADN non fonctionnel les entourant. Nous l'avons vu en 1.5.3, une proportion importante de CRMs ne satisfont pas à cette règle. Cette approche ne permet donc d'étudier que le sous-ensemble de CRMs qui a subi une forte pression de sélection depuis le dernier ancêtre commun aux espèces considérées et ne donne pas accès aux CRMs apparus récemment au sein d'une espèce.

- **Prédictions à partir de la contrainte évolutive seule**

L'alignement de séquences non-codantes orthologues fait apparaître des parties très conservées, avec peu de variations dans les séquences sous-jacentes, entourées de séquences accumulant les variations (fig. 1.22b). De telles séquences conservées sont alors interprétées comme ayant été sous sélection, les substitutions délétères ayant été rejetées au cours de l'évolution (Dermitzakis et al., 2005). Par analogie avec les empreintes à la DNAse I, on parle d'empreinte phylogénétique pour caractériser ces courtes séquences très conservées ($\sim 10\text{bp}$), traces de la fixation putative d'un facteur de transcription. Ces empreintes s'avèrent être un indicateur fiable de fonctionnalité (Kheradpour et al., 2007) et, parce qu'elles ne reposent pas sur des modèles *a priori* de fixation, elles permettent de plus de trouver des motifs de régulation non connus (Xie et al., 2005). Au niveau de séquences plus longues ($\sim 100\text{bp}$), la contrainte évolutive permet de détecter des CRMs entiers. Ainsi, comme nous l'avons vu en 1.5.3, l'utilisation de la conservation extrême permet d'atteindre 50% de taux validation (Pennacchio et al., 2006). Néanmoins, lorsque ces contraintes de conservation extrême (par exemple homme-Fugu) sont relâchées, le taux de validation tombe drastiquement, atteignant $\sim 5\%$ (Attanasio et al., 2008), montrant la nécessité d'allier le critère de conservation à d'autres données (expression, ChIP...) pour améliorer la prédition des CRMs.

- **Prédictions utilisant la phylogénie et des motifs connus**

Une approche pour améliorer les prédictions est de combiner les approches précédentes en utilisant à la fois le *clustering* en TFBS et la contrainte évolutive. À l'échelle du génome entier, cette approche permet de filtrer les résultats pour améliorer le signal de détection chez la Drosophile (Sinha et al., 2004). Du côté des mammifères, en utilisant les motifs de la base de données TRANSFAC et la conservation entre l'homme et la souris, Blanchette et al. (2006) ont créé une base de données de modules, PReMods, qui retrouve $\sim 17\%$ de CRMs connus et recoupe 40% des fragments occupés par le co-activateur et marqueur de l'activité enhancer p300. D'autres méthodes se sont concentrées sur des types cellulaires bien définis.

Chapitre 1. Introduction générale.

Par exemple, la recherche de sites conservés pour des motifs de TF des cellules sanguines connus ([Donaldson et al., 2005](#)) a permis de définir des CRMs dont 2 ont été testés et validés.

Certains efforts ont par ailleurs été menés pour sortir du cadre d'une conservation de séquence stricte en modélisant l'évolution d'un CRM fixé par un certain nombre de motifs connus. Par exemple, le modèle MorphMS ([Sinha and He, 2007](#)) cherche au sein d'un alignement de deux séquences orthologues des régions prédites par un modèle d'évolution dérivé d'un ensemble de motifs choisis par l'utilisateur. Une extension de cette approche incorpore le gain et la perte de sites de fixation, mais n'a cependant pas encore été appliquée à l'échelle du génome ([Majoros and Ohler, 2010](#)).

- **Approches utilisant la phylogénie pour générer des motifs *de novo***

De même que précédemment, tous les motifs ne sont pas connus et il peut être utile d'avoir recours à de l'apprentissage direct à partir de séquences fonctionnelles connues pour aider à la prédiction. Par exemple, l'algorithme ESPER cherche des patterns (TFBS, %GC, etc) surreprésentés dans des alignements multi-espèces de CRMs connus par rapport à des alignements d'ADN *a priori* non fonctionnel ([Taylor et al., 2006](#)). Cette méthode n'est pas restreinte à l'analyse de séquences conservées puisqu'elle peut potentiellement capturer des signatures de changements systématiques. La prédiction de régions de haut potentiel de régulation recouvre presque entièrement les prédictions de PReMods, et le test par transfection de ces régions à proximité de gènes exprimés dans les cellules érythroïdes et possédant un site pour un TF spécifique de l'érythroïde mène à un taux de validation de 50%. Une autre méthode consiste à chercher des mots surreprésentés dans un ensemble d'apprentissage de CRMs connus puis à restreindre les prédictions aux régions conservées ([Kantorovitz et al., 2009](#)). Les prédictions réalisées ont toutes été validées chez la Drosophile comme chez la souris.

1.6.3 Méthodes utilisant les marques épigénétiques et de ChIP-seq pour des TFs

- **Prédiction des promoteurs**

La méthode la plus fiable de prédiction d'un promoteur utilise le fait qu'il est toujours localisé au niveau d'un TSS, dont la position peut facilement être obtenue en alignant les séquences de l'ARN du gène correspondant sur le génome ([Trinklein et al., 2003](#)). Le taux de validation avec cette seule contrainte est très élevé : 91% ont une activité dans au moins un type cellulaire. Par ailleurs, la marque épigénétique H3K4me3 est aussi un indicateur des

promoteurs actifs dans le type cellulaire étudié (Heintzman et al., 2007) (fig. 1.14).

- **Prédiction des enhancers**

La prédiction des enhancers à partir des marques épigénétiques, comme l'acétylation des histones (Roh et al., 2005), la méthylation H3K4me1 (Heintzman et al., 2009), ou encore la présence du co-activateur p300 (Visel et al., 2009a), est très efficace, avec une expression tissu-spécifique dans ~ 80% des cas (Hardison and Taylor, 2012). Par exemple, ces différentes marques, présentes dans différents tissus, peuvent être utilisées comme autant d'entrées d'un modèle de Markov caché pour produire des prédictions fiables de CRMs tissu-spécifiques chez l'homme (Ernst et al., 2011).

En fait, les prédictions d'activité enhancer à partir de ces marques épigénétiques est plus fiable qu'en utilisant la fixation de facteurs de transcription tissu-spécifiques. Par exemple, sur 63 séquences ADN fixées *in vivo* par le facteur spécifique des cellules sanguines GATA1 chez la souris, seulement la moitié conduisent à une activité après transfection dans des cultures cellulaires (Cheng et al., 2008). Ces enhancers fonctionnels sont par ailleurs plus particulièrement associés à un site de fixation conservé pour GATA1, montrant à nouveau la nécessité de combiner les approches pour améliorer la détection. Un taux de validation similaire a été observé pour le facteur de différenciation myogénique MyoD, avec 40% de régions fixées ayant une activité après transfection en cellules.

L'utilisation de données de fixation pour plusieurs TFs à la fois semble cependant améliorer le pouvoir de prédiction. Ainsi, Tijssen et al. (2011) ont étudié la co-fixation de GATA1 avec 4 autres TFs hématopoïétiques dans des mégacaryocytes. En s'intéressant aux gènes à proximité de ces régions, ils en ont découvert plusieurs qui n'étaient pas précédemment connus comme étant important dans l'hématopoïèse. Leur fonction a été testée par knock-down, avec dans 8 cas sur les 9 testés une réduction de la production de globules rouges.

1.6.4 Validation expérimentale

Il existe plusieurs méthodes pour s'assurer de la fonctionnalité d'un CRM prédit.

Tout d'abord, une méthode indirecte donnant du crédit à la prédiction d'un CRM est d'examiner le motif d'expression du gène dont le TSS est le plus proche. Si cette expression reproduit les caractéristiques utilisées pour prédire le CRM (par exemple, s'exprimer dans le muscle pour une prédiction de CRMs utilisant l'abondance de sites de liaison de TFs muscu-

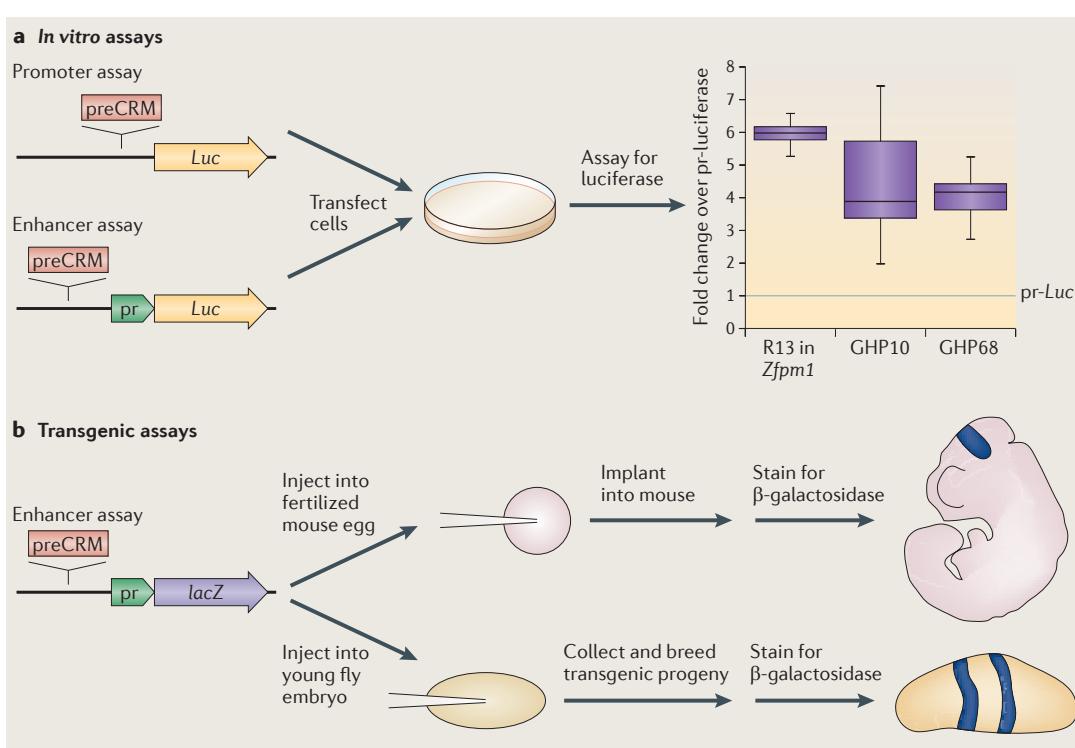


FIGURE 1.23 – Méthodes de validation des CRMs par transfection et transgenèse.

Figure tirée de ([Hartwell et al., 1999](#)) présentant les méthodes *in vitro* et *in vivo* de validation des CRMs. La région dont on souhaite tester l'activité est insérée dans un plasmide codant pour un gène rapporteur qui est transféré dans une culture cellulaire (transfection, panel a) ou dans un organisme entier (transgenèse, panel b). Dans le cas du test d'un promoteur, le CRM est placé directement en amont du gène rapporteur (on utilise généralement la luciférase *Luc*), alors que dans le cas d'un enhancer, le CRM est placé en amont d'un promoteur minimal de faible activité. L'activité de la luciférase donne une information quantitative sur l'activité de la région testée (boîtes à moustache, panel a). Dans le cas d'une transgenèse, le gène rapporteur généralement utilisé est *lacZ* qui encode la β -galactosidase. La révélation par coloration permet de visualiser en bleu les tissus au sein lesquels l'enhancer est actif.

laires), alors cela soutient l'idée (mais ne la démontre pas) que la présence du CRM en est la cause.

Une méthode plus directe permettant de démontrer qu'un fragment d'ADN régule l'expression génétique consiste en une expérience de gain de fonction dans laquelle un plasmide contenant le CRM prédict à proximité d'un gène rapporteur est introduit par transfection *in vitro* en cellule, permettant un suivi quantitatif de l'activité, ou par transgenèse *in vivo* dans

un organisme, auquel cas le suivi est plus qualitatif mais permet d'établir la spécificité spatio-temporelle (tissu et stade de développement) de l'élément de régulation (fig. 1.23). Ce type d'expérience montre que le CRM prédict est *suffisant* pour reproduire le motif génétique observé. De manière optimale, il faudrait aussi montrer par délétion ciblée de l'élément de régulation au sein du génome que ce dernier est *nécessaire* à l'expression du gène endogène.

1.6.5 Implication des CRMs dans les maladies humaines



Au cours des dernières décennies, de nombreuses mutations dans les régions codantes des gènes, impliquant des défauts structurels des protéines associées, ont pu être associées à des maladies génétiques. À l'inverse, le rôle des mutations affectant des régions non codantes n'a

Chapitre 1. Introduction générale.

été que peu exploré, essentiellement du fait de la difficulté d'annoter ces régions correctement afin de définir celles qui pourraient avoir une fonction d'intérêt. Plusieurs études ont cependant pu montrer que des variations affectant des enhancers distaux pouvaient conduire à des pathologies ([Visel et al., 2009b](#)).

L'une de ces études concerne l'enhancer spécifique du membre de *Shh* (fig. 1.24). Cet enhancer, initialement décrit chez la souris, se situe à environ 1 Mb de distance de *Shh*, au sein de l'intron d'un gène voisin. Le séquençage de cet enhancer chez plusieurs individus humains a permis d'associer une douzaine de variations mono-nucléotidiques à la polydactylie pré axiale, c'est-à-dire la présence de doigts ou d'orteils supplémentaires ([Lettice et al., 2003](#)). Des études supplémentaires chez la souris ont montré que les variations de séquences observées dans cet enhancer conduisent à une expression ectopique dans la partie antérieure du membre au cours du développement, ce qui est consistant avec la présence de doigts supplémentaires ([Masuya et al., 2007](#)). Par ailleurs, la délétion de l'enhancer orthologue de la souris entraîne la troncation des membres ([Sagai et al., 2005](#)).

Ainsi, ces résultats montrent l'importance de l'identification des enhancers pour permettre à des études de génétique humaine d'explorer le rôle potentiellement pathologique de mutations dans des régions non codantes fonctionnelles.

1.7 Bases de données

La biologie moderne est caractérisée par l'accumulation de données biologiques qu'il s'agit d'intégrer puis d'interpréter : on parle de biologie intégrative. En particulier, depuis le séquençage du génome humain il y a maintenant plus de dix ans (Lander et al., 2001), le nombre de génome séquencés n'a cessé d'augmenter, tandis que dans le même temps le prix du séquençage diminuait drastiquement (fig. 1.25). Afin de permettre la gestion et l'utilisation de ces données, de nombreux outils et bases de données ont été mis à disposition (Wasserman and Sandelin, 2004). Nous évoquons ici ceux qui nous paraissent essentiels du point de vue de la régulation en *cis*.

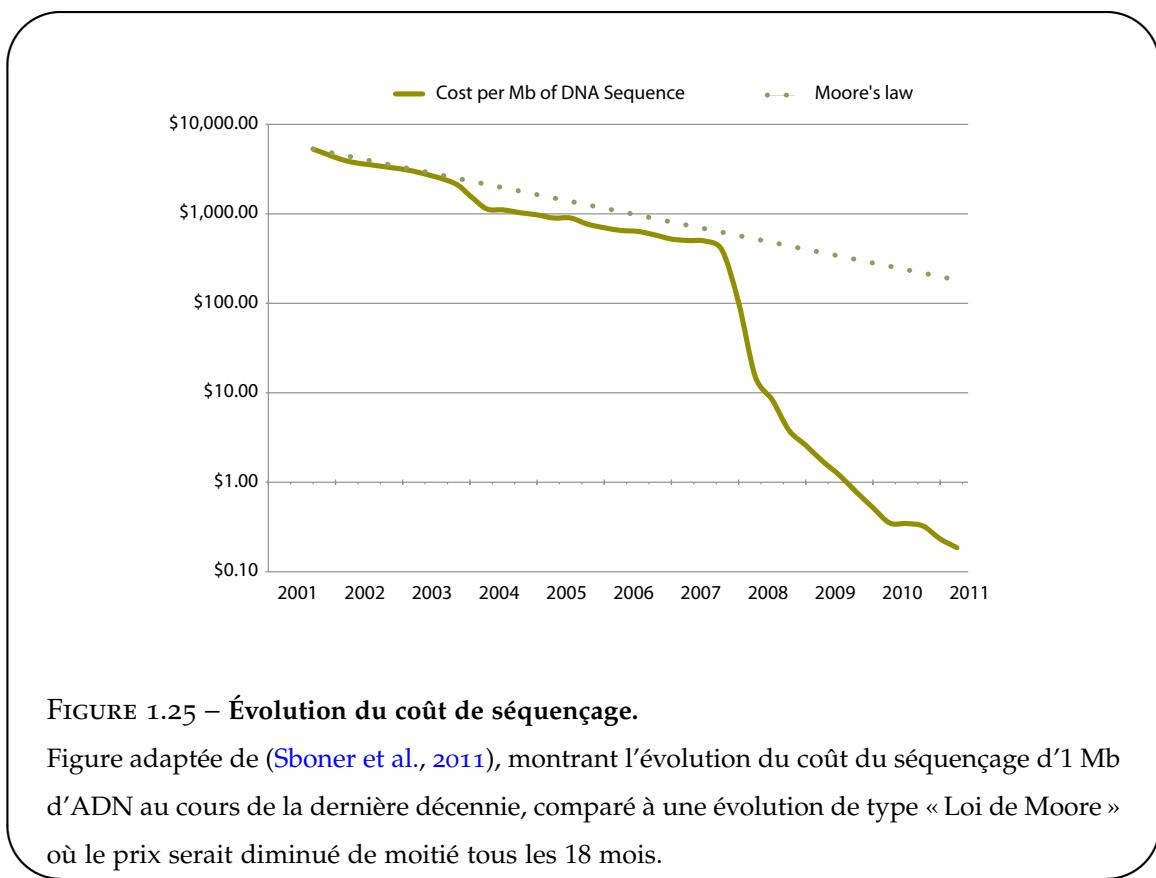


FIGURE 1.25 – Évolution du coût de séquençage.

Figure adaptée de (Sboner et al., 2011), montrant l'évolution du coût du séquençage d'1 Mb d'ADN au cours de la dernière décennie, comparé à une évolution de type « Loi de Moore » où le prix serait diminué de moitié tous les 18 mois.

1.7.1 Obtention de données génomiques

Tout d'abord, les différents génomes séquencés sont à disposition sur des bases de données publiques d'où ils peuvent être téléchargés puis analysés en aval. Parmi les plus généralistes se trouvent la base de donnée de UCSC (UCSC Genome Browser, <http://genome.ucsc.edu>)

Chapitre 1. Introduction générale.

et celle de l'EMBL (Ensembl, <http://www.ensembl.org>)³.

Sont à disposition les génomes des différentes espèces séquencées pour les différents assemblages réalisés, des alignements des génomes de différentes espèces deux par deux (*pair-wise alignments*) ou par groupes d'espèces (*multiple alignments*), ainsi qu'un certain nombre d'annotations essentielles à l'analyse de ces génomes : coordonnées des gènes (TSSs, exons, introns avec potentiellement différents transcrits alternatifs), miRNA ou lincRNA, ontologies associées, coordonnées des séquences répétitives (les *repeats*, en partie liés aux éléments transposables abordés en 1.5.3, et qui sont abondants dans les génomes vertébrés), différentes données ChIP-seq, indices de conservation⁴...

Au final, ces différentes données constituent une base de travail fiable et régulièrement mise à jour. Afin de faciliter leur obtention, il est possible d'utiliser le navigateur de tables de UCSC⁵ ou la section BioMart d'Ensembl⁶.

Situé plus en amont, le projet Galaxy (<http://galaxyproject.org>) permet à l'utilisateur de récupérer des données depuis les différentes banques existantes, puis de leur faire subir divers traitements et analyses par divers outils de bioinformatique. Cet outil, qui peut être utilisé sur internet ou bien localement, a l'avantage de permettre la sauvegarde de plans de travail ou *workflows*, successions de commandes utilisées pour traiter une entrée donnée par différents outils stéréotypés et obtenir directement le résultat final, favorisant une approche conviviale orientée utilisateur.

En guise d'exemple, nous montrons en figure 1.26 des statistiques obtenues aisément à partir d'annotations génétiques présentes sur UCSC et traitées avec Galaxy. Ces statistiques sont les distribution de tailles des régions intergéniques et introniques chez plusieurs espèces : la bactérie *Escherichia coli*, la levure *Saccharomyces cerevisiae*, le ver *Caenorhabditis elegans*, la mouche *Drosophila melanogaster*, la souris, le poulet et l'homme.

3. Les données sont accessibles sur les pages de téléchargement, respectivement <http://hgdownload.cse.ucsc.edu/downloads.html> pour UCSC et <http://www.ensembl.org/info/data/ftp/index.html> pour Ensembl

4. Pour le cas de l'assemblage mm9 de la souris, ces annotations sont accessibles à l'adresse suivante : <http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/>

5. <http://genome.ucsc.edu/cgi-bin/hgTables>

6. <http://www.ensembl.org/biomart/martview>

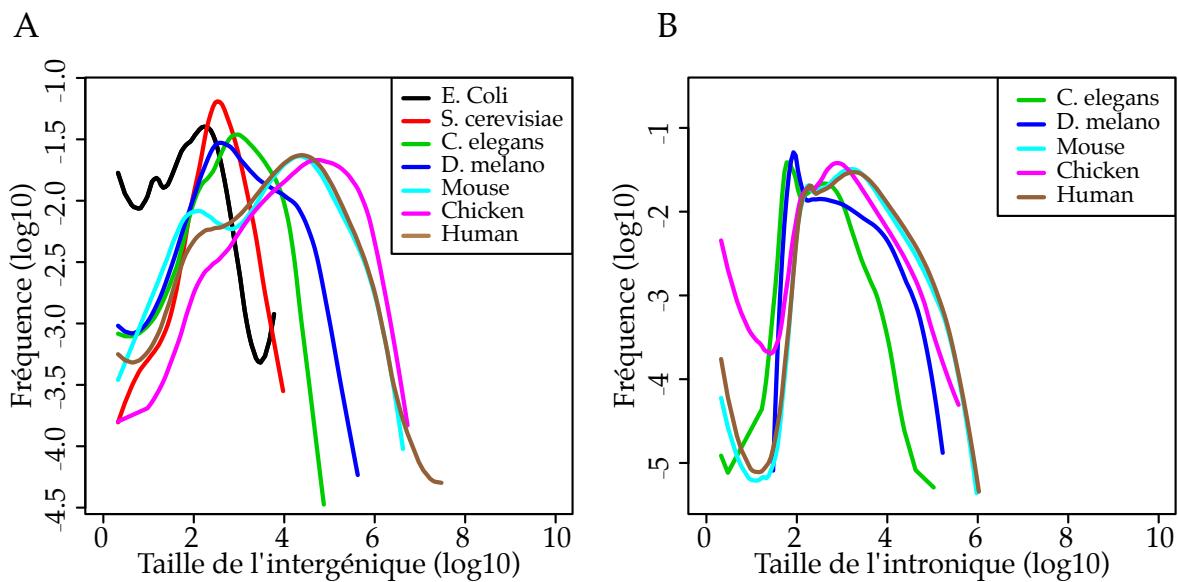


FIGURE 1.26 – Distribution des tailles intergéniques et introniques chez différentes espèces.

Distributions log-log de la taille des régions intergéniques (A) et introniques (B) chez différentes espèces. Les histogrammes sont réalisés avec un intervalle de 0.05 puis lissés avec l'estimateur local LOESS de paramètre $span = 0.3$ (logiciel R). (A) Les régions intergéniques sont définies comme les régions complémentaires aux régions transcrtes (données UCSC), celles-ci étant préalablement fusionnées pour éviter les redondances liées aux multiples transcrits d'un même gène. De la bactérie à l'homme, on observe une inflation de la quantité de génome non codant. (B) Les régions introniques sont définies par le fait qu'elles sont entourées par deux exons d'un même gène. Pour pouvoir être épissés lors de la maturation des preARNm, les introns doivent posséder des sites d'épissage, imposant une borne inférieure à leur taille pour que l'ARNm final soit fonctionnel.

1.7.2 Obtention de données sur les TFs

Nous l'avons vu, les données de fixation des TFs (ChIP-seq, ChIP-on-chip) peuvent être obtenues à partir du site UCSC Genome Browser. Ces données sont aussi généralement accessibles sur le site du NCBI (Gene Expression Omnibus, <http://www.ncbi.nlm.nih.gov/geo/>) via un numéro d'accès donné lors de la publication des données.

De nombreux modèles de TFs ont déjà été bâties préalablement à l'avènement des données haut-débit de type ChIP-seq, par exemple avec des données SELEX, et il existe des bases de données stockant les PWMs correspondantes : JAPSAR, base de données publique⁷, et TRANSFAC, qui marche par abonnement⁸. Il est à noter que ces PWMs ayant souvent été

7. <http://jaspar.cgb.ki.se>

8. <http://www.gene-regulation.com/pub/databases.html>

Chapitre 1. Introduction générale.

construites à partir d'un faible nombre de sites de fixations et de données *in vitro*, elles peuvent être relativement inadaptées à l'analyse de données *in vivo*.

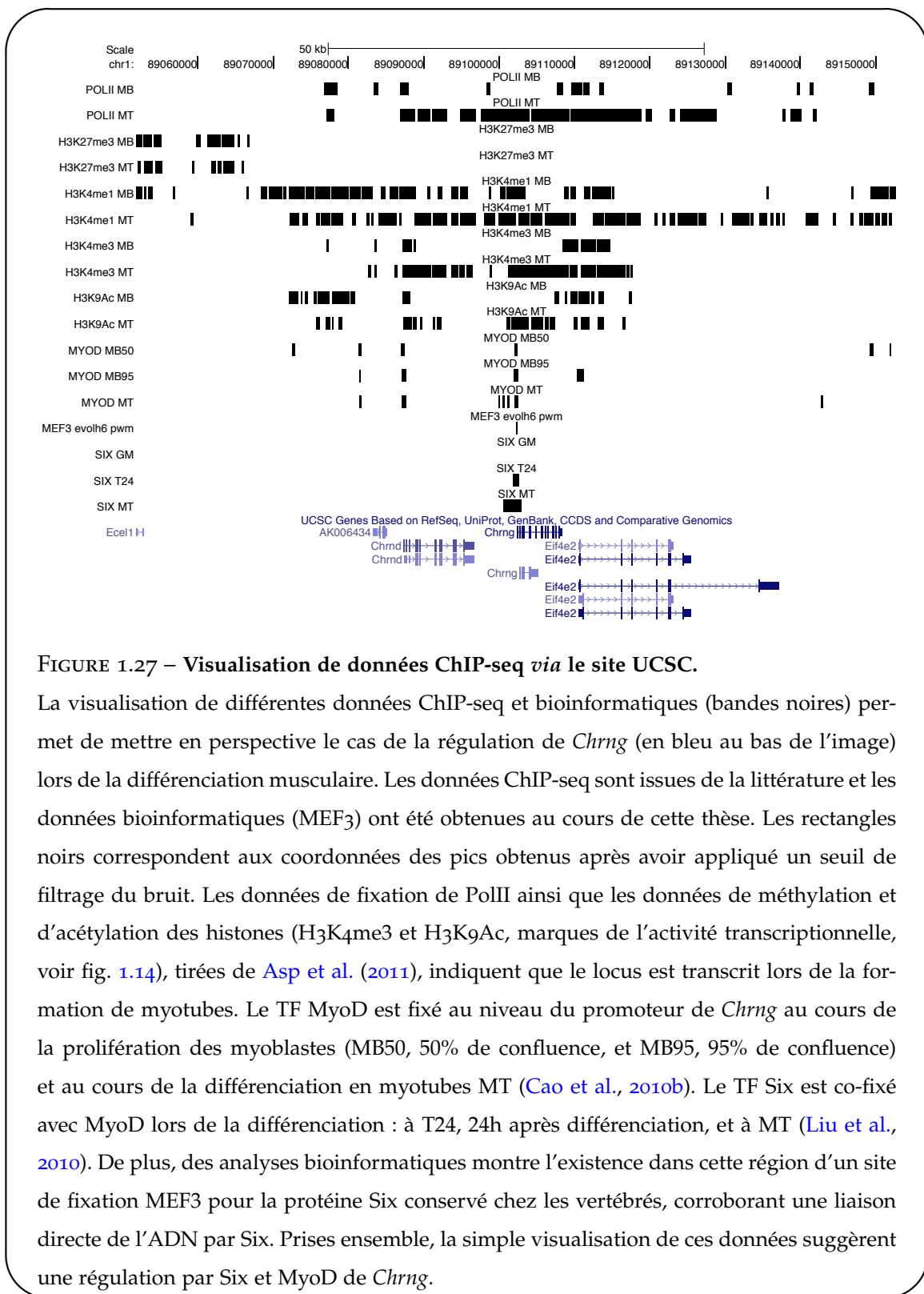
1.7.3 Outils de visualisation

Afin d'avoir une idée plus claire des événements de régulation qui se déroulent à un locus donné, il existe plusieurs outils de visualisation des annotations génomiques et épigénétiques, que ce soit sur le site du NCBI (<http://www.ncbi.nlm.nih.gov/gene>), sur Ensembl ou sur UCSC Genome Browser. Ce dernier possède notamment l'avantage qu'il est possible d'importer des données personnelles sous un grand nombre de formats, obtenues à partir de la littérature ou à partir de ses propres travaux. Ainsi, nous présentons en figure 1.27 quelques données de ChIP-seq pour des TFs musculaires et pour des marques épigénétiques, ainsi que des prédictions bioinformatiques de sites de fixation conservés pour les homéoprotéines Six réalisée par nos soins. La visualisation sur UCSC Genome Browser permet de rapidement déterminer le mode de régulation putatif du gène *Chrng* : fixation de Six et MyoD au niveau du promoteur et apparition de marques épigénétiques H3K4me1 et H3Ac sur les histones au cours de la différenciation de progéniteurs musculaires.

Par ailleurs, il existe un outil de visualisation complémentaire de ceux cités : le visualisateur de régions conservées au cours de l'évolution ECR Browser (<http://ecrbrowser.dcode.org>), intégrant de nombreux outils bioinformatiques (Loots and Ovcharenko, 2005). Ce navigateur permet de visualiser la conservation génomique d'un locus donné chez plusieurs espèces plus ou moins lointaines (par exemple souris, homme, vache, grenouille et poisson zèbre) afin de cibler l'étude de la régulation sur des régions extrêmement conservées. Il est ensuite possible d'analyser les séquences ultraconservées sélectionnées en utilisant les motifs de la base de donnée TRANSFAC via l'outil rVISTA (Loots and Ovcharenko, 2004). Un exemple d'utilisation de cet outil est donné par la découverte de plusieurs régions de régulation fonctionnelles de l'homéoprotéine Six1 possédant une extrême conservation (Sato et al., 2012).

1.7.4 Le projet ENCODE

Le projet ENCODE (pour *Encyclopedia of DNA Elements*) est un consortium de groupes de recherche internationaux financés par le NHGRI (*National Human Genome Research Institute*)



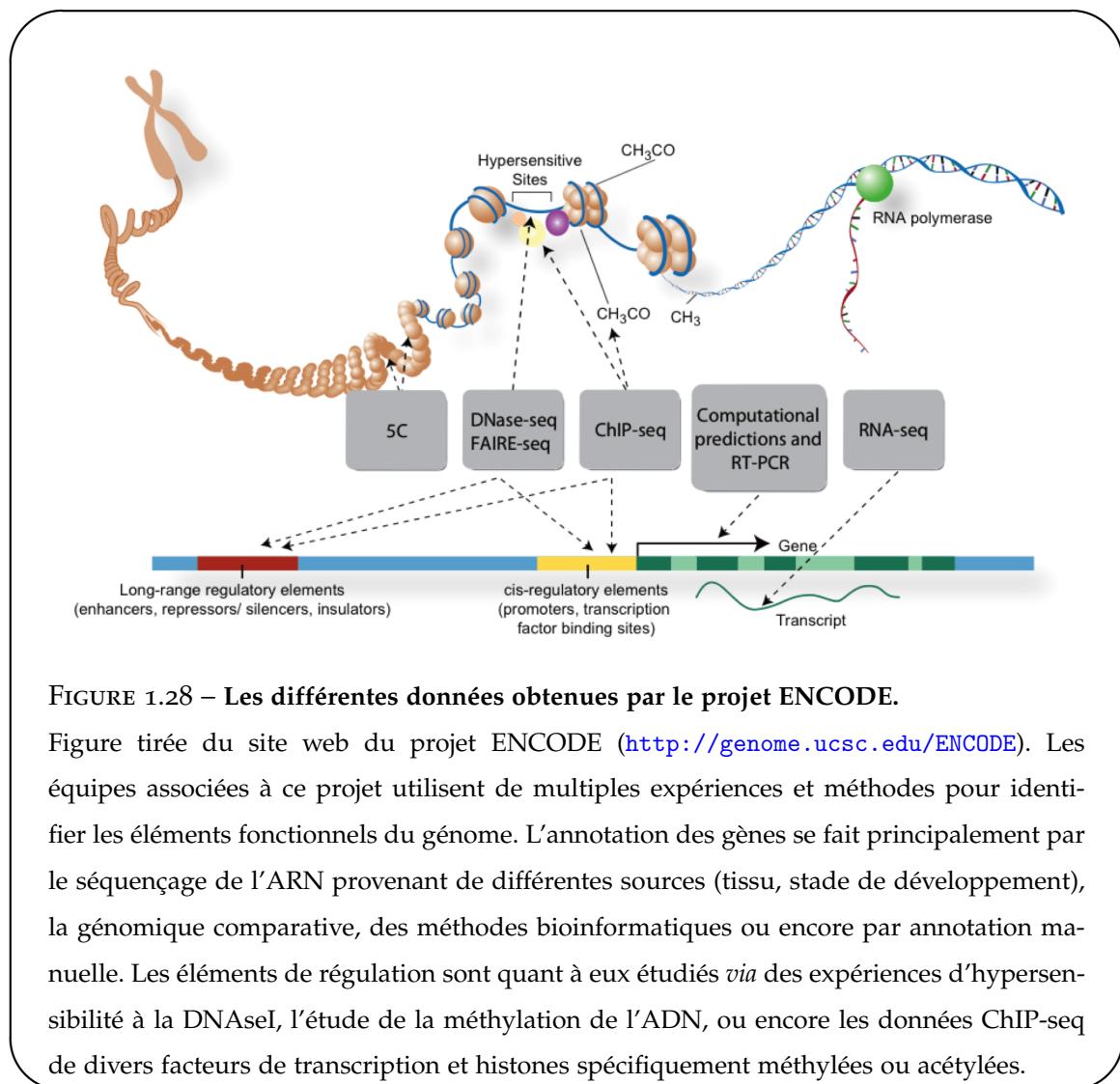


FIGURE 1.28 – Les différentes données obtenues par le projet ENCODE.

Figure tirée du site web du projet ENCODE (<http://genome.ucsc.edu/ENCODE>). Les équipes associées à ce projet utilisent de multiples expériences et méthodes pour identifier les éléments fonctionnels du génome. L'annotation des gènes se fait principalement par le séquençage de l'ARN provenant de différentes sources (tissu, stade de développement), la génomique comparative, des méthodes bioinformatiques ou encore par annotation manuelle. Les éléments de régulation sont quant à eux étudiés *via* des expériences d'hypersensibilité à la DNaseI, l'étude de la méthylation de l'ADN, ou encore les données ChIP-seq de divers facteurs de transcription et histones spécifiquement méthylées ou acétylées.

qui a vu le jour afin de systématiser les méthodes permettant l'annotation des génomes et de faciliter l'intégration des nombreuses données obtenues. Son but est de construire une liste exhaustive des éléments fonctionnels du génome humain, qu'ils agissent au niveau de l'ADN, de l'ARN ou des protéines, et des éléments de régulation qui contrôlent l'état cellulaire et l'activité des gènes. Les données sont mises à disposition du public gratuitement sur internet (<http://genome.ucsc.edu/ENCODE/>). À noter que des projets équivalents existent pour d'autres organismes, comme la souris (<http://mouseencode.org>), ou encore le ver *Caenorhabditis elegans* et la mouche *Drosophila melanogaster* (<http://www.modencode.org>).

Totalisant en septembre 2012 plus de 1600 expériences dans plus de 147 types cellulaires, les premières conclusions pointent vers une profusion d'événements de régulation, loin de

L'idée d'ADN poubelle (*junk DNA*) : ainsi, 80% du génome est associé à un événement biochimique associé à de la formation d'ARN ou au remodelage de la chromatine, $\sim 400,000$ régions possèdent un état chromatinien caractéristique des enhancers et $\sim 70,000$ des promoteurs (ENCODE Project Consortium et al., 2012). Depuis mai 2013, les données ChIP-seq de 161 TFs couvrant 91 types cellulaires ont été mises à disposition sur UCSC Genome Browser⁹.

9. <http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeAwgTfbsUniform>

Chapitre 2

Modèles de fixation des Facteurs de Transcription à l'ADN.

3.1 77

Introduction du chapitre 2

intro : insister sur description de ce qui s'est fait ensuite : ne pas traduire l'article mais approfondir les points non abordés (entropie maximale, info directe etc)

- L'énergie de fixation. Les Facteurs de Transcription peuvent s'accrocher à l'ADN. La fixation est décrite par une énergie qui peut se décomposer en deux composantes. L'une est indépendante de la séquence et prend en considération la courbure de l'ADN etc. L'autre dépend de la séquence. Cette dernière peut être décrite par divers modèles de fixation.
- Description des modèles existants.
- Différentes données biologiques utilisées : PBM, SELEX, ChIP.
- Différences in vitro et in vivo.

2.1 Les modèles de fixation

2.1.1 Modèles de maximum d'entropie

La théorie de l'information offre un cadre conceptuel permettant de déterminer les probabilités d'un ensemble d'états étant données plusieurs contraintes mesurables, ou *observables*. L'étape clé consiste à maximiser une fonctionnelle connue sous le nom d'entropie (Jaynes, 1957; Shannon, 1948) sur l'ensemble des distributions de probabilités des états étant données les contraintes imposées. Cette fonctionnelle s'écrit (Sigal et al., 2006)

$$S[P_m] = - \sum_{\{s\}} P_m(s) \ln P_m(s) \quad (2.1)$$

où $P_m(s)$ est la probabilité modèle d'une séquence d'ADN s appartenant à l'ensemble $\{s\}$ des sites de fixation d'un facteur de transcription. Notons $\mathcal{O}_\alpha(s)$ une quantité attachée à s . Dans notre cas, cette quantité peut représenter la présence d'un certain nucléotide à une position donnée, ou d'une paire de nucléotide à deux positions données. Ce que l'on nomme observable correspond en fait à la moyenne de cette quantité sur l'ensemble des états donnés : $\langle \mathcal{O}_\alpha(s) \rangle_r$, où l'indice r signifie que nous moyennons en utilisant la statistique P_r sur les séquences observées. La contrainte associée s'écrit :

$$\langle \mathcal{O}_\alpha(s) \rangle_m = \langle \mathcal{O}_\alpha(s) \rangle_r \quad (2.2)$$

où l'indice m signifie que la moyenne est prise sur la distribution modèle. Nous pouvons alors écrire le Lagrangien suivant

$$\mathcal{L} = - \sum_{\{s\}} P(s) \ln P(s) + \lambda \left(\sum_{\{s\}} P(s) - 1 \right) + \sum_\alpha \beta_\alpha (\langle \mathcal{O}_\alpha(s) \rangle_m - \langle \mathcal{O}_\alpha(s) \rangle_r) \quad (2.3)$$

où λ et les β_α sont les multiplicateurs de Lagrange correspondant respectivement à la contrainte de normalisation de la distribution de probabilité et aux différentes observables \mathcal{O}_α . La maximisation de ce Lagrangien est obtenue en annulant la dérivée fonctionnelle par rapport à la distribution de probabilité P_m :

$$\frac{\delta \mathcal{L}}{\delta P_m(s)} = 0 = -\ln P_m(s) - 1 + \lambda + \sum_\alpha \beta_\alpha \mathcal{O}_\alpha(s) \quad (2.4)$$

La solution peut finalement se mettre sous la forme

$$P_m(s) = \frac{1}{Z} e^{-\mathcal{H}(s)} \quad (2.5)$$

où \mathcal{H} est l'Hamiltonien du système :

$$\mathcal{H} = \sum_{\alpha} \beta_{\alpha} \mathcal{O}_{\alpha}(s) \quad (2.6)$$

et Z est la fonction de partition permettant la normalisation de la distribution P_m :

$$Z = \sum_{\{s\}} \exp[-\mathcal{H}(s)]. \quad (2.7)$$

- **Le modèle PWM**
- **Le modèle de corrélation de paires**

Fixation de jauge.

2.1.2 Modèles de mélange

2.2 Description des données biologiques

2.2.1 Les données ChIP

Les données que nous utilisons proviennent d'expériences ChIP-on-chip réalisées chez la mouche (*Drosophila melanogaster*) et d'expériences ChIP-seq réalisées chez la souris (*Mus musculus*). Ces données ont été récupérées à partir de la littérature (Zinzen et al., 2009; Chen et al., 2008) et à partir des données du projet ENCODE (ENCODE Project Consortium, 2011) accessibles à partir du site internet de UCSC¹⁰, pour un total de 27 Facteurs de Transcription. Parmi eux, il y a 5 Facteurs de Transcription impliqués dans le développement de la mouche : Bap, Bin, Mef2, Tin, Twi, 11 Facteurs de Transcription régulant les cellules souches chez les mammifères : c-Myc, E2f1, Esrrb, Klf4, Nanog, n-Myc, Oct4, Sox2, Stat3, Tcfcp2l1, Zfx, et 11 facteurs impliqués dans la myogenèse chez les mammifères : Cebpb, E2f4, Fosl1, Max, MyoD, Myog, Nrsf, Smad1, Srf, Tcf3, Usf1. Au total, il y a entre 678 et 38292 pics de ChIP, avec une taille moyenne de 280bp.

¹⁰. <http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeCaltechTfbs/>

Les séquences d'ADN peuvent contenir un certain nombre de séquences « polluantes » peu informatives issues de rétrotransposons ou de duplication excessives de dinucléotides. Ces séquences répétées, ou *repeats*, sont en grand nombre et peuvent donc biaiser la statistique lors de la recherche de sites de fixation. Pour éviter ce biais, ces séquences ont été masquées à l'aide du logiciel RepeatMasker (Smit et al., 1996-2010).

2.2.2 Statistique « background » des séquences

Présence de corrélations.

2.3 Présentation de l'algorithme

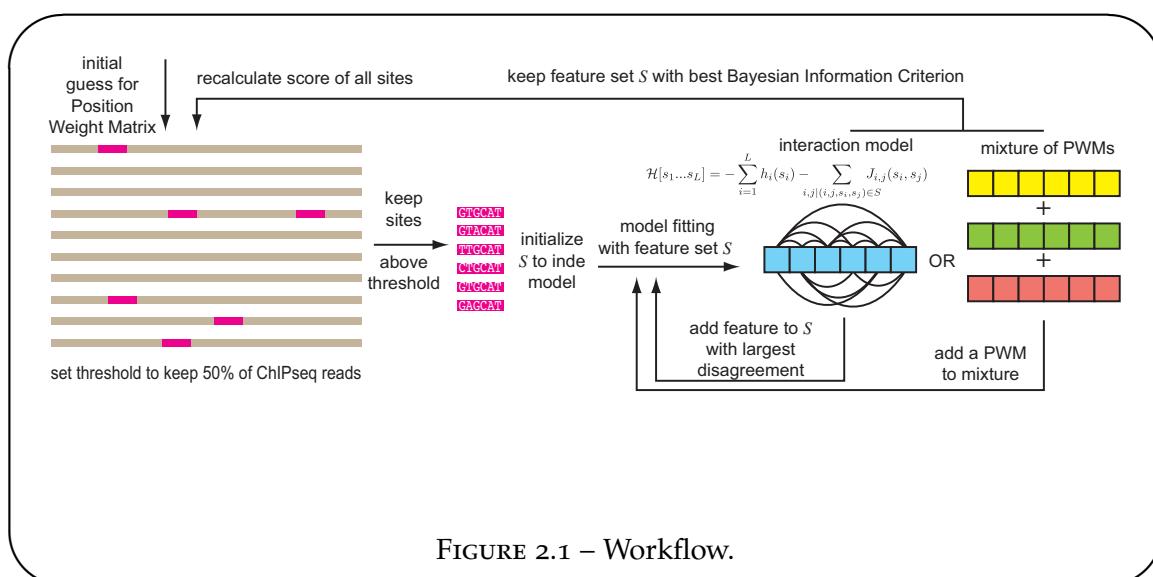


FIGURE 2.1 – Workflow.

Descente de gradient.

2.4 Performance des modèles

2.5 Analyse des corrélations

2.5.1 Quantification par l'Information Directe

2.5.2 Description par des patterns de Hopfield

2.6 Comparaison avec des données *in vitro*

2.6.1 Conclusion de la section 2.6

Chapitre 3

Imogene : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle

5.1	85
-----	-------	----

Introduction du chapitre 3

- Trouver des motifs d'ADN sans *a priori*.
- Grammaire des enhancers : rigidité ou flexibilité.

3.1

Chapitre 4

Étude de la différenciation épidermale chez la drosophile

Introduction du chapitre 4

4.1

Conclusion du chapitre 4

Chapitre 5

Étude de la différenciation musculaire chez la souris

Introduction du chapitre 5

idees : décrire interface UCSC ncRNA dissection des enhancers pour comprendre la logique des enhancers

5.1

Conclusion du chapitre 5

Conclusion

RÉSUMÉ

PERSPECTIVES

Conclusion

Bibliographie

- Aerts, S. (2012). Chapter 5 - Computational Strategies for the Genome-Wide Identification of cis-Regulatory Elements and Transcriptional Targets. *Current Topics in Developmental Biology : Transcriptional Switches During Development*, 98 :121–145. (Page 41.)
- Alon, U. (2007a). An Introduction to Systems Biology : Design Principles of Biological Circuits (Mathematical and Computational Biology Series vol 10). (Page 17.)
- Alon, U. (2007b). Network motifs : theory and experimental approaches. *Nat Rev Genet*, 8(6) :450–461. (Page 17.)
- Asp, P., Blum, R., Vethantham, V., Parisi, F., Micsinai, M., Cheng, J., Bowman, C., Kluger, Y., and Dynlacht, B. D. (2011). PNAS Plus : Genome-wide remodeling of the epigenetic landscape during myogenic differentiation. *Proceedings of the National Academy of Sciences*, pages 1–11. (Page 63.)
- Attanasio, C., Reymond, A., Humbert, R., Lyle, R., Kuehn, M. S., Neph, S., Sabo, P. J., Goldy, J., Weaver, M., Haydock, A., Lee, K., Dorschner, M., Dermitzakis, E. T., Antonarakis, S. E., and Stamatoyannopoulos, J. A. (2008). Assaying the regulatory potential of mammalian conserved non-coding sequences in human cells. *Genome Biol*, 9(12) :R168. (Page 53.)
- Aurell, E., d'Hérouël, A., Malmnäs, C., and Vergassola, M. (2007). Transcription factor concentrations versus binding site affinities in the yeast *S. cerevisiae*. *Physical biology*, 4 :134. (Page 27.)
- Bartel, D. P. (2009). MicroRNAs : target recognition and regulatory functions. *Cell*, 136(2) :215–33. (Page 14.)
- Baylies, M. K., Bate, M., and Ruiz Gomez, M. (1998). Myogenesis : a view from Drosophila. *Cell*, 93(6) :921–7. (Page 19.)
- Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98(3) :387–96. (Page 41.)
- Berg, O. and von Hippel, P. (1987). Selection of DNA binding sites by regulatory proteins :

Bibliographie

- Statistical-mechanical theory and application to operators and promoters. *Journal of molecular biology*, 193(4) :723–743. (Page 21.)
- Berg, O. G., Winter, R. B., and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, 20(24) :6929–48. (Page 21.)
- Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., 3rd, and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11) :1429–35. (Page 29.)
- Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., and Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc Natl Acad Sci U S A*, 99(2) :757–62. (Page 50.)
- Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev*, 16(1) :6–21. (Page 12.)
- Blackwell, T. K. and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science*, 250(4984) :1104–10. (Page 30.)
- Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganière, J., Lefèvre, C., Deblois, G., Giguère, V., Ferretti, V., Bergeron, D., Coulombe, B., and Robert, F. (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res*, 16(5) :656–68. (Page 53.)
- Blau, H. M., Pavlath, G. K., Hardeman, E. C., Chiu, C. P., Silberstein, L., Webster, S. G., Miller, S. C., and Webster, C. (1985). Plasticity of the differentiated state. *Science*, 230(4727) :758–66. (Page 9.)
- Bolouri, H. and Davidson, E. H. (2002). Modeling DNA sequence-based cis-regulatory gene networks. *Dev Biol*, 246(1) :2–13. (Page 40.)
- Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H. H., and Liu, E. T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*, 18(11) :1752–1762. (Page 45.)

- Brazma, A., Parkinson, H., Schlitt, T., and Shojatalab, M. (2001). A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays. http://www.ebi.ac.uk/microarray/biology_intro.html. (Page 5.)
- Buchler, N. E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9) :5136–41. (Page 40.)
- Campbell, C. T. and Kim, G. (2007). SPR microscopy and its applications to high-throughput analyses of biomolecular binding events and their kinetics. *Biomaterials*, 28(15) :2380–92. (Page 29.)
- Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G. J., Parker, M. H., Macquarrie, K. L., Davison, J., Morgan, M. T., Ruzzo, W. L., Gentleman, R. C., and Tapscott, S. J. (2010a). Genome-wide MyoD binding in skeletal muscle cells : a potential for broad cellular reprogramming. *Developmental Cell*, 18(4) :662–74. (Page 52.)
- Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G. J., Parker, M. H., Macquarrie, K. L., Davison, J., Morgan, M. T., Ruzzo, W. L., Gentleman, R. C., and Tapscott, S. J. (2010b). Genome-wide MyoD binding in skeletal muscle cells : a potential for broad cellular reprogramming. *Developmental Cell*, 18(4) :662–74. (Page 63.)
- Carlson, C. D., Warren, C. L., Hauschild, K. E., Ozers, M. S., Qadir, N., Bhimsaria, D., Lee, Y., Cerrina, F., and Ansari, A. Z. (2010). Specificity landscapes of DNA binding molecules elucidate biological function. *Proc Natl Acad Sci U S A*, 107(10) :4544–9. (Page 30.)
- Carninci, P., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6) :626–35. (Page 39.)
- Carvajal, J. J., Keith, A., and Rigby, P. W. J. (2008). Global transcriptional regulation of the locus encoding the skeletal muscle determination genes Mrf4 and Myf5. *Genes & development*, 22(2) :265–76. (Pages 39 et 40.)
- Chen, X., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6) :1106–17. (Page 70.)
- Cheng, Y., King, D. C., Dore, L. C., Zhang, X., Zhou, Y., Zhang, Y., Dorman, C., Abebe, D., Kumar, S. A., Chiaromonte, F., Miller, W., Green, R. D., Weiss, M. J., and Hardison, R. C.

Bibliographie

- (2008). Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif. *Genome Res*, 18(12) :1896–905. (Page 55.)
- Chung, J. H., Whiteley, M., and Felsenfeld, G. (1993). A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, 74(3) :505–14. (Page 41.)
- Cordaux, R. and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10) :691–703. (Page 47.)
- Davis, R. L., Weintraub, H., and Lassar, A. B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell*, 51(6) :987–1000. (Page 9.)
- Dermitzakis, E. T. and Clark, A. G. (2002). Evolution of transcription factor binding sites in Mammalian gene regulatory regions : conservation and turnover. *Mol Biol Evol*, 19(7) :1114–21. (Page 45.)
- Dermitzakis, E. T., Reymond, A., and Antonarakis, S. E. (2005). Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet*, 6(2) :151–7. (Page 53.)
- Djordjevic, M., Sengupta, A. M., and Shraiman, B. I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Res*, 13(11) :2381–90. (Page 24.)
- Donaldson, I. J., Chapman, M., Kinston, S., Landry, J. R., Knezevic, K., Piltz, S., Buckley, N., Green, A. R., and Göttgens, B. (2005). Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. *Hum Mol Genet*, 14(5) :595–601. (Page 54.)
- ENCODE Project Consortium (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *Plos Biol*, 9(4) :e1001046. (Page 70.)
- ENCODE Project Consortium, et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414) :57–74. (Page 65.)
- Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shores, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E.

- (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345) :43–9. (Page 55.)
- Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*, 9(5) :397–405. (Page 47.)
- Fields, D. S., He, Y., Al-Uzri, A. Y., and Stormo, G. D. (1997). Quantitative specificity of the Mnt repressor. *J Mol Biol*, 271(2) :178–94. (Page 30.)
- Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., and Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, pages 1–5. (Page 48.)
- Furusawa, C. and Kaneko, K. (2012). A Dynamical-Systems View of Stem Cell Biology. *Science*, 338(6104) :215–217. (Page 6.)
- Gerland, U., Moroz, J., and Hwa, T. (2002). Physical constraints and functional characteristics of transcription factor–DNA interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 99(19) :12015. (Pages 21, 24, 25 et 26.)
- Giocomo, L. M., Moser, M.-B., and Moser, E. I. (2011). Computational models of grid cells. *Neuron*, 71(4) :589–603. (Page 23.)
- Graf, T. and Enver, T. (2009). Forcing cells to change lineages. *Nature*, 462(7273) :587–94. (Page 9.)
- Greer, E. L. and Shi, Y. (2012). Histone methylation : a dynamic mark in health, disease and inheritance. *Nat Rev Genet*, 13(5) :343–57. (Page 14.)
- Gurdon, J. B. and Melton, D. A. (2008). Nuclear reprogramming in cells. *Science*, 322(5909) :1811–5. (Page 10.)
- Hammond, S. M., Caudy, A. A., and Hannon, G. J. (2001). Post-transcriptional gene silencing by double-stranded RNA. *Nat Rev Genet*, 2(2) :110–9. (Page 14.)
- Hannon, G. J. (2002). RNA interference. *Nature*, 418(6894) :244–51. (Page 14.)
- Hardison, R. C. and Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics*, 13(7) :469–483. (Pages 38 et 55.)

Bibliographie

- Hartwell, L., Hopfield, J., Leibler, S., and Murray, A. (1999). From molecular to modular cell biology. *Nature*, 402(6761) :47. (Pages [51](#) et [56](#).)
- Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, 39(3) :311–8. (Page [55](#).)
- Heintzman, N. D., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243) :108–12. (Pages [40](#) et [55](#).)
- Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods*, 6(4) :283–9. (Page [36](#).)
- Hong, J.-W., Hendrix, D. A., and Levine, M. S. (2008). Shadow enhancers as a source of evolutionary novelty. *Science*, 321(5894) :1314. (Page [47](#).)
- Jaynes, E. (1957). Information theory and statistical mechanics. II. *Physical review*, 108(2) :171. (Page [69](#).)
- Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E., and Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*, 20(6) :861–73. (Page [31](#).)
- Jolma, A., Yan, J., Whittington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013). DNA-Binding Specificities of Human Transcription Factors. *Cell*, 152(1-2) :327–39. (Page [31](#).)
- Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J., and Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314) :430–435. (Page [48](#).)

- Kantorovitz, M. R., Kazemian, M., Kinston, S., Miranda-Saavedra, D., Zhu, Q., Robinson, G. E., Göttgens, B., Halfon, M. S., and Sinha, S. (2009). Motif-blind, genome-wide discovery of cis-regulatory modules in *Drosophila* and mouse. *Developmental Cell*, 17(4) :568–79. (Pages 52 et 54.)
- Kaufmann, S. (1993). The origins of order. (Page 8.)
- Keim, C. N., Martins, J. L., Abreu, F., Rosado, A. S., de Barros, H. L., Borojevic, R., Lins, U., and Farina, M. (2004). Multicellular life cycle of magnetotactic prokaryotes. *FEMS Microbiol Lett*, 240(2) :203–8. (Page 6.)
- Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res*, 17(12) :1919–31. (Page 53.)
- Kinney, J. B., Tkacik, G., and Callan, C. G. (2007). Precise physical models of protein-DNA interaction from high-throughput data. *Proc Natl Acad Sci USA*, 104(2) :501–6. (Page 29.)
- Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., and Kageyama, Y. (2010). Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science*, 329(5989) :336–9. (Page 14.)
- Kulessa, H., Frampton, J., and Graf, T. (1995). GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes Dev*, 9(10) :1250–62. (Page 9.)
- Kulkarni, M. M. and Arnosti, D. N. (2003). Information display by transcriptional enhancers. *Development*, 130(26) :6569–75. (Pages 42 et 43.)
- Lander, E. S., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921. (Page 59.)
- Lässig, M. (2007). From biophysics to evolutionary genetics : statistical aspects of gene regulation. *BMC Bioinformatics*, 8(Suppl 6) :S7. (Pages 20 et 25.)
- Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., and Simon, I. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594) :799. (Pages 16 et 17.)
- Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in

Bibliographie

- the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*, 12(14) :1725–35. (Pages [39](#), [57](#) et [58](#).)
- Lberman, L. M. and Stathopoulos, A. (2009). Design flexibility in cis-regulatory control of gene expression : Synthetic and comparative evidence. *Developmental Biology*, 327(2) :578–589. (Pages [44](#) et [45](#).)
- Liu, Y., Chu, A., Chakroun, I., Islam, U., and Blais, A. (2010). Cooperation between myogenic regulatory factors and SIX family transcription factors is important for myoblast differentiation (SI). *Nucleic acids research*. (Page [63](#).)
- Liu, Y.-H., Jakobsen, J. S., Valentin, G., Amarantos, I., Gilmour, D. T., and Furlong, E. E. M. (2009). A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development. *Developmental Cell*, 16(2) :280–91. (Page [18](#).)
- Loots, G. G. and Ovcharenko, I. (2004). rVISTA 2.0 : evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*, 32(Web Server issue) :W217–21. (Page [62](#).)
- Loots, G. G. and Ovcharenko, I. (2005). Dcode.org anthology of comparative genomic tools. *Nucleic Acids Res*, 33(Web Server issue) :W56–64. (Page [62](#).)
- Ludwig, M. Z., Bergman, C., Patel, N. H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403(6769) :564–7. (Page [45](#).)
- Maerkl, S. and Quake, S. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809) :233. (Page [28](#).)
- Majoros, W. H. and Ohler, U. (2010). Modeling the evolution of regulatory elements by simultaneous detection and alignment with phylogenetic pair HMMs. *PLoS Comput Biol*, 6(12) :e1001037. (Page [54](#).)
- Maniatis, T., Goodbourn, S., and Fischer, J. A. (1987). Regulation of inducible and tissue-specific gene expression. *Science*, 236(4806) :1237–45. (Page [39](#).)
- Masuya, H., Sezutsu, H., Sakuraba, Y., Sagai, T., Hosoya, M., Kaneda, H., Miura, I., Kobayashi, K., Sumiyama, K., Shimizu, A., Nagano, J., Yokoyama, H., Kaneko, S., Sakurai, N., Okagaki, Y., Noda, T., Wakana, S., Gondo, Y., and Shiroishi, T. (2007). A series of ENU-induced

- single-base substitutions in a long-range cis-element altering Sonic hedgehog expression in the developing mouse limb bud. *Genomics*, 89(2) :207–14. (Page 58.)
- McGregor, A., Orgogozo, V., Delon, I., Zanet, J., Srinivasan, D., Payre, F., and Stern, D. (2007). Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature*, 448(7153) :587–590. (Page 47.)
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs : simple building blocks of complex networks. *Science*, 298(5594) :824–7. (Page 17.)
- Moses, A. M., Pollard, D. A., Nix, D. A., Iyer, V. N., Li, X.-Y., Biggin, M. D., and Eisen, M. B. (2006). Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol*, 2(10) :e130. (Page 45.)
- Nagaraj, V. H., O'flanagan, R. A., and Sengupta, A. M. (2008). Better estimation of protein-DNA interaction parameters improve prediction of functional sites. *BMC Biotechnol*, 8(1) :94. (Pages 30 et 31.)
- Nurse, P. and Hayles, J. (2011). The Cell in an Era of Systems Biology. *Cell*. (Page 10.)
- Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K., and Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*, 39(6) :730–2. (Page 45.)
- Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Rinaldi, N. J., Murray, H. L., Volkert, T. L., Schreiber, J., Rolfe, P. A., Gifford, D. K., Fraenkel, E., Bell, G. I., and Young, R. A. (2004). Control of pancreas and liver gene expression by HNF transcription factors. *Science*, 303(5662) :1378–81. (Pages 16 et 17.)
- Oliphant, A. R., Brandl, C. J., and Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides : analysis of yeast GCN4 protein. *Mol Cell Biol*, 9(7) :2944–9. (Page 30.)
- Ondek, B., Gloss, L., and Herr, W. (1988). The SV40 enhancer contains two distinct levels of organization. *Nature*, 333(6168) :40–5. (Page 39.)
- Panne, D. (2008). The enhanceosome. *Curr Opin Struct Biol*, 18(2) :236–42. (Page 43.)

Bibliographie

Park, P. J. (2009). ChIP-seq : advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10) :669–80. (Page [35](#).)

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B. L., Couronne, O., Eisen, M. B., Visel, A., and Rubin, E. M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118) :499–502. (Pages [43](#) et [53](#).)

Perry, M. W., Boettiger, A. N., and Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the *Drosophila* embryo. *Proc Natl Acad Sci U S A*, 108(33) :13570–5. (Page [48](#).)

Phillips, J. E. and Corces, V. G. (2009). CTCF : master weaver of the genome. *Cell*, 137(7) :1194–211. (Page [41](#).)

Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E. M., Couronne, O., and Pennacchio, L. A. (2006). Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res*, 16(7) :855–63. (Page [43](#).)

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E. D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics*, 3 :30. (Page [50](#).)

Recillas-Targa, F., Pikaart, M. J., Burgess-Beusse, B., Bell, A. C., Litt, M. D., West, A. G., Gaszner, M., and Felsenfeld, G. (2002). Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proc Natl Acad Sci U S A*, 99(10) :6883–8. (Page [41](#).)

Roh, T.-Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev*, 19(5) :542–52. (Page [55](#).)

Roulet, E., Busso, S., Camargo, A. A., Simpson, A. J. G., Mermod, N., and Bucher, P. (2002). High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol*, 20(8) :831–5. (Page [30](#).)

- Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development*, 132(4) :797–803. (Pages 57 et 58.)
- Sato, S., Ikeda, K., Shioi, G., Nakao, K., Yajima, H., and Kawakami, K. (2012). Regulation of Six1 expression by evolutionarily conserved enhancers in tetrapods. *Dev Biol*, 368(1) :95–108. (Page 62.)
- Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., and Gerstein, M. B. (2011). The real cost of sequencing : higher than you think ! *Genome Biol*, 12(8) :125. (Page 59.)
- Schirm, S., Jiricny, J., and Schaffner, W. (1987). The SV40 enhancer can be dissected into multiple segments, each with a different cell type specificity. *Genes Dev*, 1(1) :65–74. (Page 39.)
- Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, Â., Kutter, C., Brown, G. D., Marshall, A., Flückeck, P., and Odom, D. T. (2012). Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell*, 148(1-2) :335–348. (Page 47.)
- Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flückeck, P., and Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding SI. *Science*, 328(5981) :1036–40. (Page 45.)
- Schoborg, T. A. and Labrador, M. (2010). The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. *J Mol Evol*, 70(1) :74–84. (Page 41.)
- Schones, D. E. and Zhao, K. (2008). Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet*, 9(3) :179–91. (Page 15.)
- Schroeder, M. D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E. D., and Gaul, U. (2004). Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol*, 2(9) :E271. (Page 50.)
- Shannon, C. (1948). A Mathematical Theory of Communication. *Bell Syst Tech J*, 27(4) :623–656. (Page 69.)

Bibliographie

- Shen-Orr, S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature genetics*, 31(1) :64–68. (Page 17.)
- Shumaker-Parry, J. S., Aebersold, R., and Campbell, C. T. (2004). Parallel, quantitative measurement of protein binding to a 120-element double-stranded DNA array in real time using surface plasmon resonance microscopy. *Anal Chem*, 76(7) :2071–82. (Page 29.)
- Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y., Rosenfeld, N., Danon, T., Perzov, N., and Alon, U. (2006). Variability and memory of protein levels in human cells. *Nature*, 444(7119) :643–646. (Page 69.)
- Sinha, S. and He, X. (2007). MORPH : probabilistic alignment combined with hidden Markov models of cis-regulatory modules. *PLoS Comput Biol*, 3(11) :e216. (Page 54.)
- Sinha, S., Schroeder, M. D., Unnerstall, U., Gaul, U., and Siggia, E. D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in *Drosophila*. *BMC Bioinformatics*, 5 :129. (Page 53.)
- Slutsky, M. and Mirny, L. A. (2004). Kinetics of protein-DNA interaction : facilitated target location in sequence-dependent potential. *Biophys J*, 87(6) :4021–35. (Page 21.)
- Smit, A. F. A., Hubley, R., and Green, P. (1996-2010). RepeatMasker Open-3.0. <http://www.repeatmasker.org>. (Page 71.)
- Smith, A. D., Sumazin, P., Xuan, Z., and Zhang, M. Q. (2006). DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci U S A*, 103(16) :6275–80. (Page 52.)
- Smith, A. D., Sumazin, P., and Zhang, M. Q. (2005). Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A*, 102(5) :1560–5. (Page 52.)
- Stormo, G. and Fields, D. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends in biochemical sciences*, 23(3) :109–113. (Page 22.)
- Stormo, G. D. and Zhao, Y. (2007). Putting numbers on the network connections. *Bioessays*, 29(8) :717–21. (Page 28.)

- Stormo, G. D. and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. *Nature Reviews Genetics*, 11(11) :751–60. (Pages [27](#) et [32](#).)
- Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4) :663–76. (Page [10](#).)
- Taylor, J., Tyekucheva, S., King, D. C., Hardison, R. C., Miller, W., and Chiaromonte, F. (2006). ESPERR : learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res*, 16(12) :1596–604. (Page [54](#).)
- Thurman, R. E., et al. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414) :75–82. (Page [37](#))
- Tijssen, M. R., Cvejic, A., Joshi, A., Hannah, R. L., Ferreira, R., Forrai, A., Bellissimo, D. C., Oram, S. H., Smethurst, P. A., Wilson, N. K., Wang, X., Ottersbach, K., Stemple, D. L., Green, A. R., Ouwehand, W. H., and Göttgens, B. (2011). Genome-wide analysis of simultaneous GATA_{1/2}, RUNX₁, FLI₁, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell*, 20(5) :597–609. (Page [55](#).)
- Tirosh, I., Weinberger, A., Bezalel, D., Kaganovich, M., and Barkai, N. (2008). On the relation between promoter divergence and gene expression evolution. *Mol Syst Biol*, 4 :159. (Page [45](#).)
- Trinklein, N. D., Aldred, S. J. F., Saldanha, A. J., and Myers, R. M. (2003). Identification and functional analysis of human transcriptional promoters. *Genome Res*, 13(2) :308–12. (Page [54](#).)
- Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment : RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968) :505–10. (Page [30](#).)
- U.S. Department of Energy (2001). Genomes to life : accelerating biological discovery (Office of Biological and Environmental Research and Office of Advanced Scientific Computing Research of the U.S. Department of Energy). http://genomicscience.energy.gov/roadmap/GTLcomplete_web.pdf. (Pages [11](#) et [13](#).)
- Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors : function, expression and evolution. *Nat Rev Genet*, 10(4) :252–63. (Page [12](#).)

Bibliographie

- Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (2009a). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231) :854–8. (Page 55.)
- Visel, A., Rubin, E. M., and Pennacchio, L. A. (2009b). Genomic views of distant-acting enhancers. *Nature*, 461(7261) :199–205. (Pages 33, 57 et 58.)
- Waddington, C. H. et al. (1957). The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.*, pages ix+–262. (Pages 6 et 7.)
- Wallace, J. A. and Felsenfeld, G. (2007). We gather together : insulators and genome organization. *Curr Opin Genet Dev*, 17(5) :400–7. (Page 41.)
- Wang, Q., Carroll, J. S., and Brown, M. (2005). Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol Cell*, 19(5) :631–42. (Page 40.)
- Warren, C. L., Kratochvil, N. C. S., Hauschild, K. E., Foister, S., Brezinski, M. L., Dervan, P. B., Phillips, G. N., Jr, and Ansari, A. Z. (2006). Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci U S A*, 103(4) :867–72. (Page 29.)
- Wasserman, W. and Fickett, J. (1998). Identification of regulatory regions which confer muscle-specific gene expression1. *Journal of molecular biology*, 278(1) :167–181. (Page 50.)
- Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4) :276–87. (Pages 23 et 59.)
- Weintraub, H., Tapscott, S. J., Davis, R. L., Thayer, M. J., Adam, M. A., Lassar, A. B., and Miller, A. D. (1989). Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc Natl Acad Sci U S A*, 86(14) :5434–8. (Page 18.)
- Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2) :307–19. (Pages 48 et 49.)
- Wilczynski, B. and Furlong, E. E. M. (2010). Challenges for modeling global gene regulatory

networks during development : Insights from Drosophila. *Developmental Biology*, 340(2) :161–169. (Page 40.)

Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., Conboy, C. M., Vanes, L., Tybulewicz, V. L. J., Fisher, E. M. C., Tavaré, S., and Odom, D. T. (2008). Species-specific transcription in mice carrying human chromosome 21. *Science*, 322(5900) :434–8. (Page 45.)

Wilson, M. D. and Odom, D. T. (2009). Evolution of transcriptional control in mammals. *Curr Opin Genet Dev*, 19(6) :579–85. (Pages 39, 45 et 46.)

Winter, R. B., Berg, O. G., and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor-operator interaction : kinetic measurements and conclusions. *Biochemistry*, 20(24) :6961–77. (Page 21.)

Winter, R. B. and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The Escherichia coli repressor-operator interaction : equilibrium measurements. *Biochemistry*, 20(24) :6948–60. (Page 21.)

Wright, W. E., Binder, M., and Funk, W. (1991). Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. *Mol Cell Biol*, 11(8) :4104–10. (Page 30.)

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031) :338–45. (Page 53.)

Zeitlinger, J., Zinzen, R. P., Stark, A., Kellis, M., Zhang, H., Young, R. A., and Levine, M. (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev*, 21(4) :385–90. (Page 47.)

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9) :R137. (Page 34.)

Zhao, Y., Granas, D., and Stormo, G. D. (2009). Inferring binding energies from selected binding sites. *PLoS Comput Biol*, 5(12) :e1000590. (Page 24.)

Zhou, Q. and Wong, W. H. (2004). CisModule : de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A*, 101(33) :12114–9. (Page 52.)

Bibliographie

Zinzen, R., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269) :65–70. (Pages [45](#) et [70](#).)

Zykovich, A., Korf, I., and Segal, D. J. (2009). Bind-n-Seq : high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic Acids Res*, 37(22) :e151. (Page [31](#).)

Résumé

Mots-clés: Régulation génétique, Facteur de transcription, Modèle de Potts, Phylogénétique, Algorithme bayésien, différenciation musculaire, trichomes.

Abstract

Cellular differentiation and tissue specification depend in part on the establishment of specific transcriptional programs of gene expression. These programs result from the interpretation of genomic regulatory information by sequence-specific transcription factors (TFs). Decoding this information in sequenced genomes is a key issue. First, we present models that describe the interaction between the TFs and the DNA sequences they bind to, called Transcription Factor Binding Sites (TFBSs). Using a Potts model inspired from spin glass physics along with high-throughput binding data for a variety of Drosophilae and mammals TFs, we show that TFBSs exhibit correlations among nucleotides and that the account of their contribution in the binding energy greatly improves the predictability of genomic TFBSs. Then, we present a Bayesian, phylogeny-based algorithm designed to computationally identify the Cis-Regulatory Modules (CRMs) that control gene expression in a set of co-regulated genes. Starting with a small number of CRMs in a reference species as a training set, but with no a priori knowledge of the factors acting in trans, the algorithm uses the over-representation and conservation of TFBSs among related species to predict putative regulatory elements along with genomic CRMs underlying co-regulation. We show several applications of this algorithm both in Drosophila and vertebrates. We also present an extension of the algorithm to the case of pattern recognition, showing that CRMs with different patterns of expression can be distinguished on the sole basis of their DNA motifs content.

Keywords: Gene regulation, Transcription Factor, Potts Model, Phylogeny, Bayesian algorithm, muscle differentiation, trichomes.

thèse: version du lundi 1^{er} juillet 2013 à 16 h 57