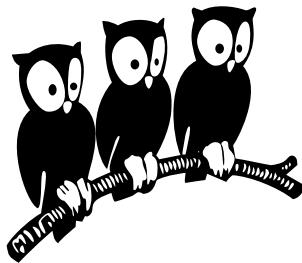


Département de Physique
École Normale Supérieure

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. ZZZ | Examinateur |
| M. ZZZ | Président du jury |
| M. ZZZ | Rapporteur |
|
 | |
| M. Pascal Maire | Membre invité |

these:version du lundi 17 juin 2013 à 17 h 53

Remerciements

...

thèse:*version du lundi 17 juin 2013 à 17 h 53*

Table des matières

Liste des figures	vii
Principales abréviations utilisées	ix
Avant-propos	1
Chapitre 1 - Introduction générale.	3
1.1 Le phénotype cellulaire	4
1.2 Les réseaux de régulation génétique	7
1.3 Modèles mathématiques des interactions protéine-ADN	14
1.4 Mesures expérimentales des interactions protéine-ADN	20
1.5 Les modules de cis-régulation	27
1.6 Banques de données	37
Chapitre 2 - Modèles de fixation des Facteurs de Transcription à l'ADN.	39
2.1 Les modèles de fixation	41
2.2 Description des données biologiques	42
2.3 Présentation de l'algorithme	42
2.4 Performance des modèles	42
2.5 Analyse des corrélations	42
2.6 Comparaison avec des données <i>in vitro</i>	42
Chapitre 3 - <i>Imogene</i> : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle	45
3.1	47
Chapitre 4 - Étude de la différenciation épidermale chez la drosophile	49
4.1	51
Chapitre 5 - Étude de la différenciation musculaire chez la souris	53
5.1	55
Conclusion	57
Résumé	67
Abstract	67

Liste des figures

Introduction générale.	3
1.1 Le paysage de la différenciation cellulaire	5
1.2 Spécification spatio-temporelle du type cellulaire	6
1.3 Différents exemples de reprogrammation cellulaire	7
1.4 Vision cybernétique du traitement de l'information par la cellule	8
1.5 Un réseau de régulation génétique type	9
1.6 Caractéristiques de l'épigénome	10
1.7 Exemples de motifs dans les réseaux de régulation génétique	12
1.8 Évolution du réseau transcriptionnel : l'exemple de la régulation myogénique.	13
1.9 Différents états du facteur de transcription	15
1.10 Construction et utilisation du modèle PWM	18
1.11 Étapes d'une expérience de ChIP-seq	24
1.12 Résolution des expériences ChIP-on-chip et ChIP-seq	25
1.13 Expérience d'empreinte à la DNase I chez la levure : vers une résolution au nucléotide près	26
1.14 Les différents types de CRMs et leurs marques épigénétiques	28
1.15 Différents CRMs conduisent à différents patterns d'expression	30
1.16 Régulation spatio-temporelle par des CRMs distincts	31
1.17 Deux modèles d'enhancers : enhanceosome et billboard	32
1.18 L'enhanceosome de l'interferon- β	33
1.21 De l'enhancer au super-enhancer	36
1.22 Méthodes de validation des CRMs par transfection et transgenèse	37
1.23 Approches pour la prédiction des CRMs	38
1.24 Évolution du coût de séquençage	38
 Modèles de fixation des Facteurs de Transcription à l'ADN.	 39
2.1 Description graphique de l'algorithme.	43
 <i>Imogene</i> : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle	 45
 Étude de la différenciation épidermale chez la drosophile	 49
 Étude de la différenciation musculaire chez la souris	 53

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>)
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 17 juin 2013 à 17 h 53

Avant-propos

Chapitre 1

Introduction générale.

1.1	Le phénotype cellulaire	4
1.1.1	Qu'est-ce que le phénotype d'une cellule?	4
1.1.2	La différenciation cellulaire	4
1.1.3	La cellule dans l'organisme : une spécification spatio-temporelle	6
1.1.4	La reprogrammation cellulaire	6
1.2	Les réseaux de régulation génétique	7
1.2.1	Vision cybernétique de la cellule	7
1.2.2	Divers modes de régulation	8
1.2.3	Câblage du réseau et fonction	11
1.2.4	Évolution des réseaux génétiques	13
1.3	Modèles mathématiques des interactions protéine-ADN	14
1.3.1	Modes de recherche du site de fixation par le TF	15
1.3.2	Modèle PWM	16
1.3.3	Modèle biophysique	17
1.3.4	Modèle thermodynamique	18
1.4	Mesures expérimentales des interactions protéine-ADN	20
1.4.1	Approches <i>in vitro</i> : MITOMI, SPR, PBM, CSI, SELEX, et HT-SELEX	20
1.4.2	Approche clonale : la technique de simple hybride	22
1.4.3	Approches <i>in vivo</i> : ChIP-on-chip, ChIP-seq, DNase I	23
1.5	Les modules de cis-régulation	27
1.5.1	Les différents types de CRMs	27
1.5.2	Encodage de patterns spatiaux	30
1.5.3	Grammaire des enhancers : enhanceosome vs billboard	31
1.5.4	Évolution des enhancers	33
1.5.5	Les « shadow enhancers »	34
1.5.6	Les super enhancers	36
1.5.7	Validation expérimentale	36
1.5.8	Prédiction des CRMs	37
1.6	Banques de données	37
1.6.1	Séquences génomiques et alignements	37
1.6.2	Annotations (TSSs, repeats...)	37
1.6.3	Jaspar et Transfac	37
1.6.4	Visualisation sur UCSC	37
1.6.5	Le projet ENCODE	37

1.1 Le phénotype cellulaire

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

Tous les organismes sont constitués de cellules de l'ordre de quelques microns, facilement observables à l'aide d'un simple microscope optique. Chaque cellule consiste en 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 qui vont nous intéresser dans cette thèse. Les cellules qui les constituent sont eucaryotes, c'est-à-dire qu'elles possèdent un noyau renfermant le matériel génétique.¹

Bien que possédant le même matériel génétique, les cellules d'un organisme apparaissent d'emblée comme hétérogènes, que ce soit dans la forme ou dans les 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.

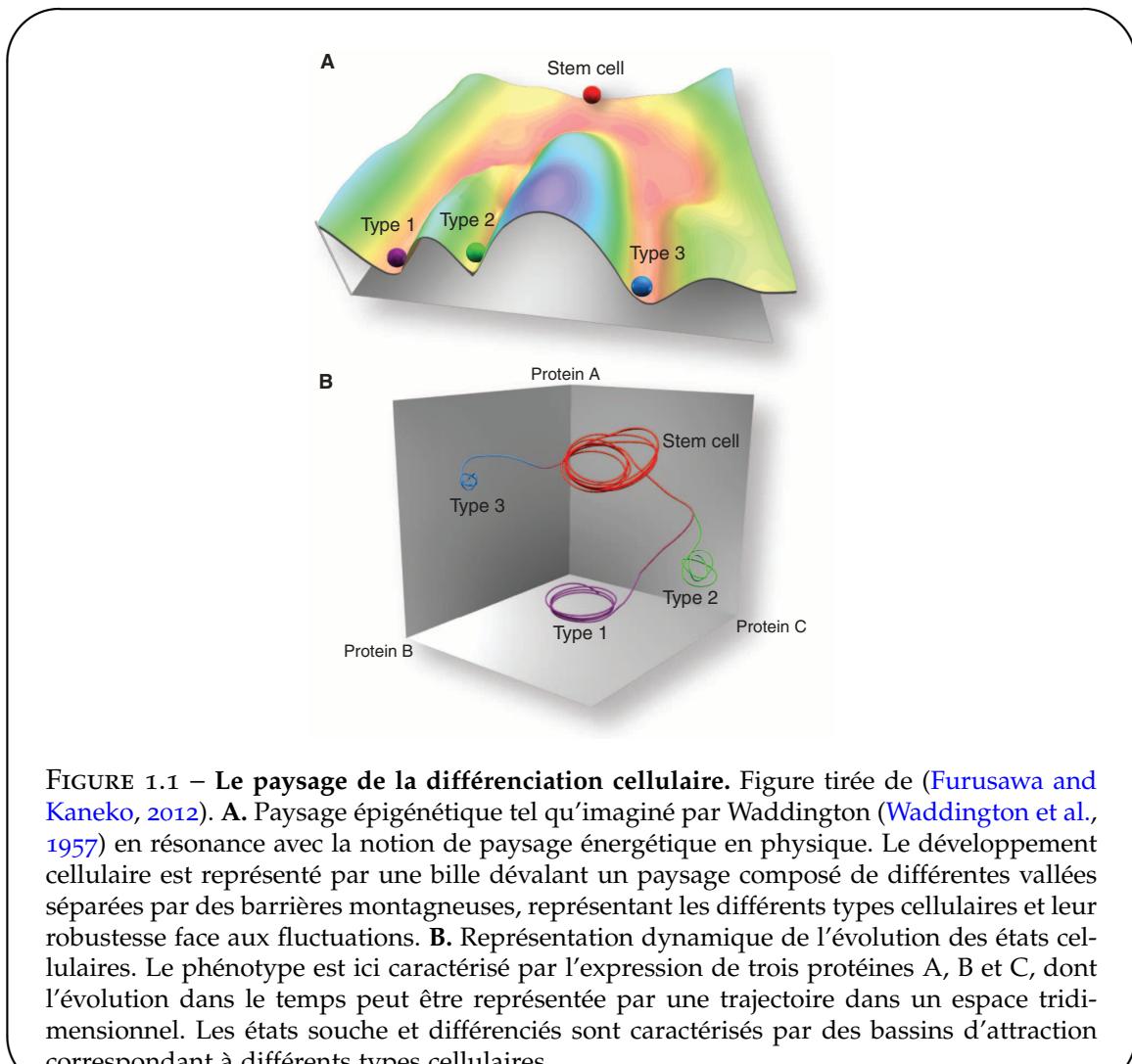
Cette diversité semble néanmoins limitée. Aussi, parmi les $\sim 6 \cdot 10^{13}$ cellules du corps humain, on peut distinguer ~ 320 différents types cellulaires (Brazma et al., 2001). Bien entendu, ce nombre dépend du seuil de similarité choisi : deux cellules d'un même type ont peu de chance d'exprimer 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 des 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 (littéralement « exhiber un type » en grec). Ce phénotype est le résultat de la modulation par des facteurs environnementaux de l'expression génétique qui détermine le contenu en protéines de la cellule.

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. Ainsi, au cours du développement d'un organisme, les cellules empruntent un chemin unidirectionnel de différenciation qui restreint peu à peu le nombre de types cellulaires qu'elles peuvent potentiellement devenir, passant d'un état souche totipotent à des états pluripotents successifs avant la différenciation finale. Ainsi, la formation des cellules somatiques, qui sont les cellules du corps n'étant ni souches ni germinales (cellules donnant naissance aux gamètes ou cellules sexuelles), est le résultat d'un processus de différenciation initial au cours duquel les cellules souches 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 ou les os (mésoderme), la peau et le système nerveux (ectoderme).

¹. Il existe cependant quelques cas connus d'organismes multicellulaires procaryotes, par exemple chez les bactéries magnétotactiques (Keim et al., 2004).



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 encore 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 manière habituelle, l'expression des protéines (et donc des gènes qui les produisent) domine ces composants, et

Chapitre 1. Introduction générale.

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, causent un changement de l'état cellulaire, ce qui se traduit par une trajectoire dans l'espace d'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 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



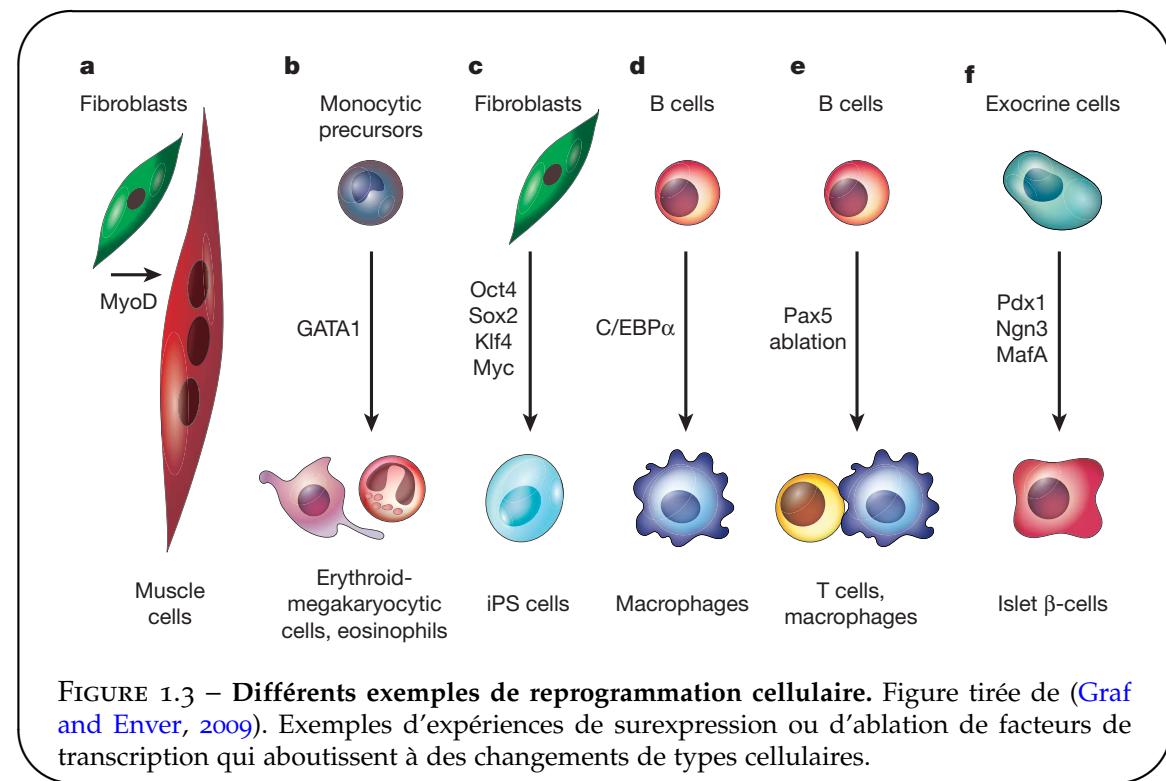
FIGURE 1.2 – Spécification spatio-temporelle du type cellulaire.

Hybridization *in situ* de *Myog*, marqueur des cellules musculaires squelettiques différenciées, chez un embryon de souris de 11.5 jours. Le pattern de spécification des cellules myogéniques est clairement visible au niveau des futures vertèbres. L'image est tirée de la base de donnée Embryos disponible sur <http://embrys.jp>.

Un fait remarquable à propos de la différenciation cellulaire est que celle-ci 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 – et du temps – les étapes de développement se succédant de manière irréversible –. Ainsi, la différenciation des cellules observe certains *patterns* spatio-temporels bien définis : par exemple, dans le cas de la formation des muscles, le marqueur des cellules du muscle squelettique *Myog* est exprimé chez la souris dès 8 jours embryonnaires au niveau des somites, les futures vertèbres de la souris adulte (voir fig 1.2).

1.1.4 La reprogrammation cellulaire

Dans les paragraphes précédents, nous avons présenté la vision classique selon laquelle des cellules souches totipotentes se différencient en des cellules de moins en moins plastiques, jusqu'à atteindre un état différencié stable. Néanmoins, depuis plusieurs décénies, différentes expériences ont exhibé la plasticité des états différenciés. Par exemple, Blau et al. ont montré en 1985 que des programmes d'expression génétiques dormants peuvent être exprimés de manière dominante dans des cellules différenciées par la fusion de différents types cellulaires ([Blau et al., 1985](#)). Puis différents travaux ont montré qu'il était possible de convertir



des lignées de cellules en introduisant certaines protéines régulatrices de la transcription, ou Facteurs de Transcription (TFs) ([Davis et al., 1987; Kulessa et al., 1995](#)) (voir fig 1.3). Parallèlement, des expériences réalisées chez plusieurs espèces ont montré que le transfert de noyaux de cellules différencierées embryonnaires ou adultes dans un oeuf énucleé 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 ([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 est gènes est en effet régulée par des protéines, les facteurs de transcription, qui sont elles-mêmes exprimées par d'autres gènes, créant ainsi des interactions entre gènes. Par ailleurs, les protéines peuvent réguler l'activité d'autres protéines, et certains ARN issus de la transcription de gènes non codants opèrent aussi de manière primordiale dans la régulation de l'activité génétique, le tout formant un réseau complexe

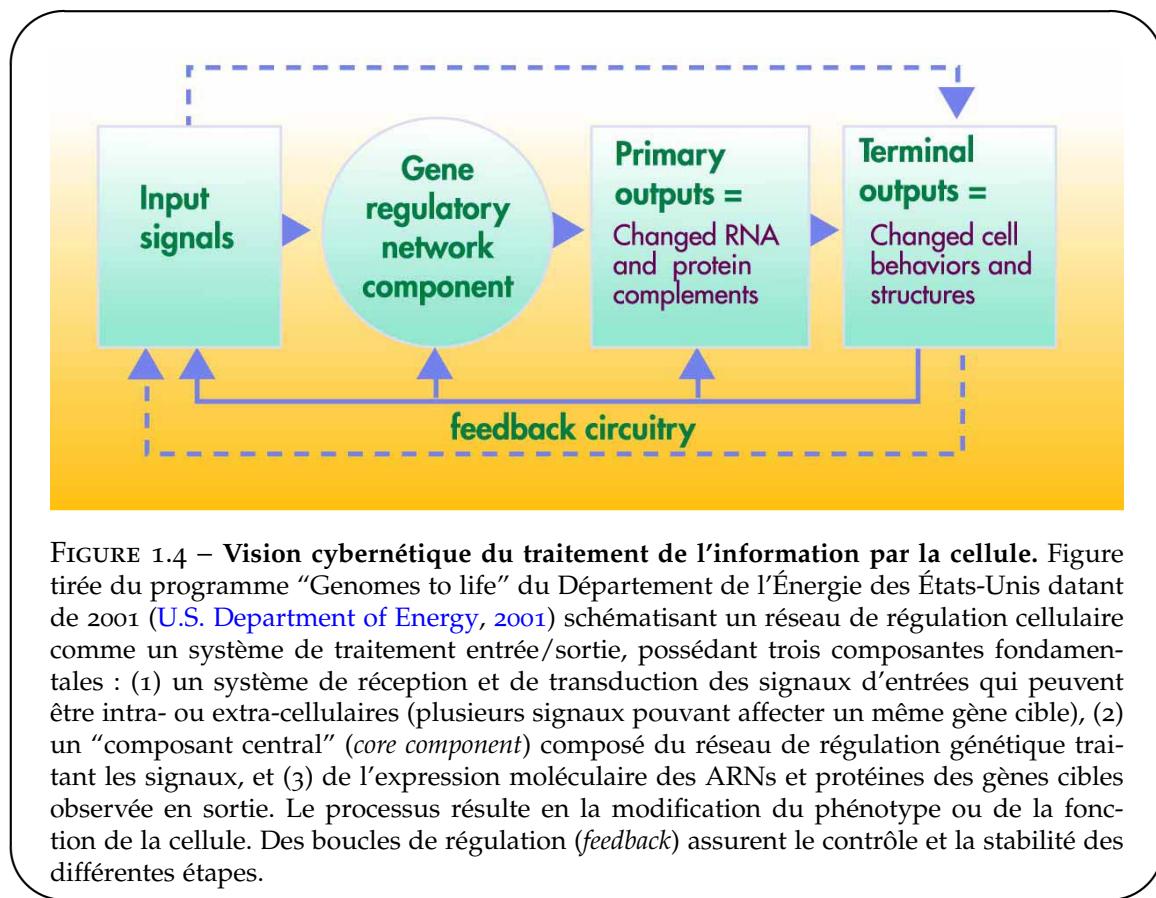


FIGURE 1.4 – Vision cybernétique du traitement de l’information par la cellule. Figure tirée du programme “Genomes to life” du Département de l’Énergie des États-Unis datant de 2001 ([U.S. Department of Energy, 2001](#)) schématisant un réseau de régulation cellulaire comme un système de traitement entrée/sortie, 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) de l’expression moléculaire des ARNs et protéines des gènes cibles observée en sortie. Le processus résulte en la modification du phénotype ou de la fonction de la cellule. Des boucles de régulation (*feedback*) assurent le contrôle et la stabilité des différentes étapes.

d’interactions. La compréhension de ce réseau et des fonctions qu’il englobe forme le socle de la discipline de biologie des systèmes. Dans ce cadre, la cellule est vue comme une machine interprétant différents signaux reçus en entrée et qui, une fois traités par le réseau interne de régulation, 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 et de changer d’état sont nombreux. Nous allons nous concentrer ici sur ceux internes aux réseaux génétiques, et affectant au final la production de protéines ou d’ARNs et donc l’état cellulaire (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 ([Vazquez 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

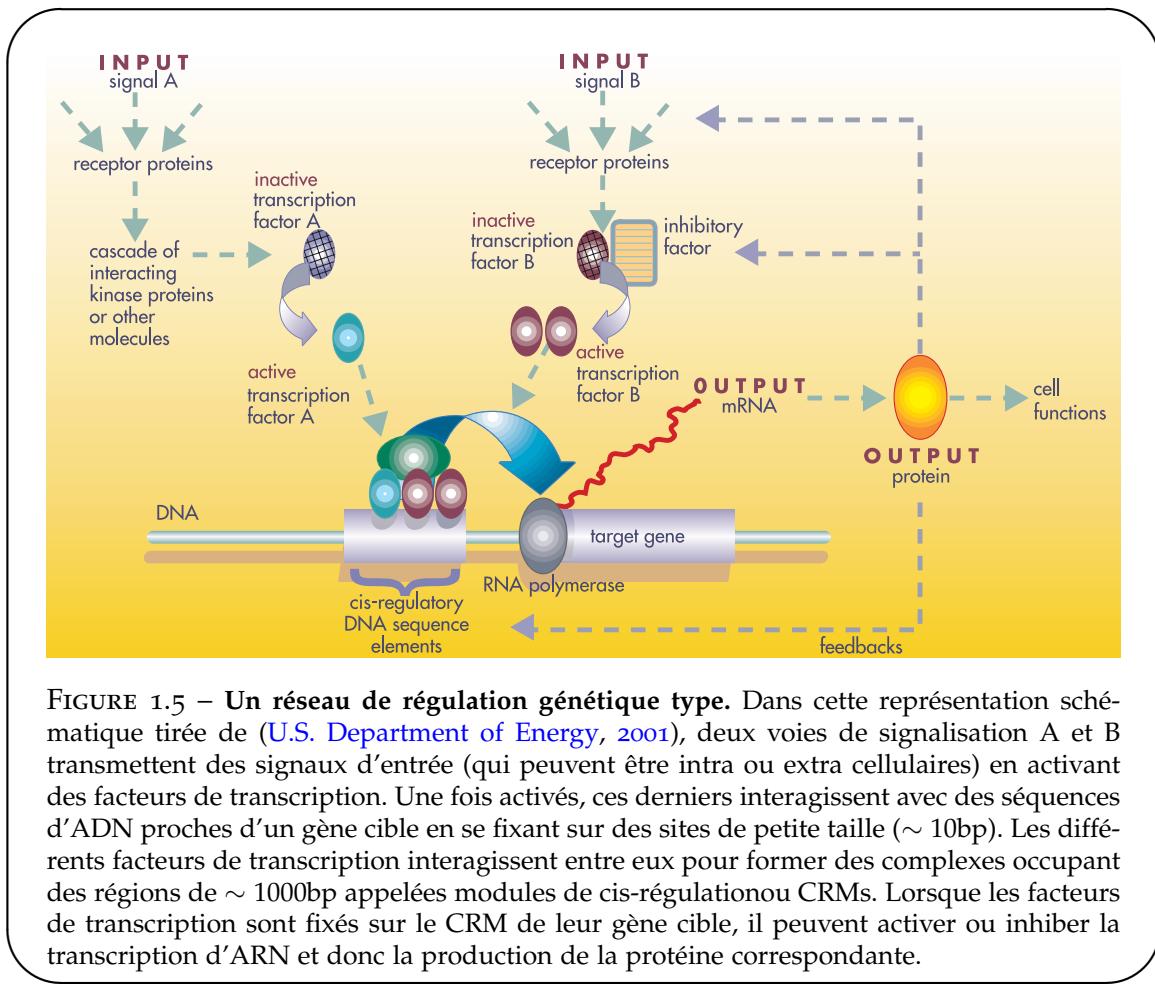


FIGURE 1.5 – Un réseau de régulation génétique type. Dans cette représentation schématique tirée de (U.S. Department of Energy, 2001), deux voies de signalisation A et B transmettent des signaux d'entrée (qui peuvent être intra ou extra cellulaires) en activant des facteurs de transcription. Une fois 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. 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.

sur l'ADN de $\sim 10\text{bp}$ 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 $\sim 1000\text{bp}$ 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 : on parle de 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. Ainsi, la méthy-

Chapitre 1. Introduction générale.

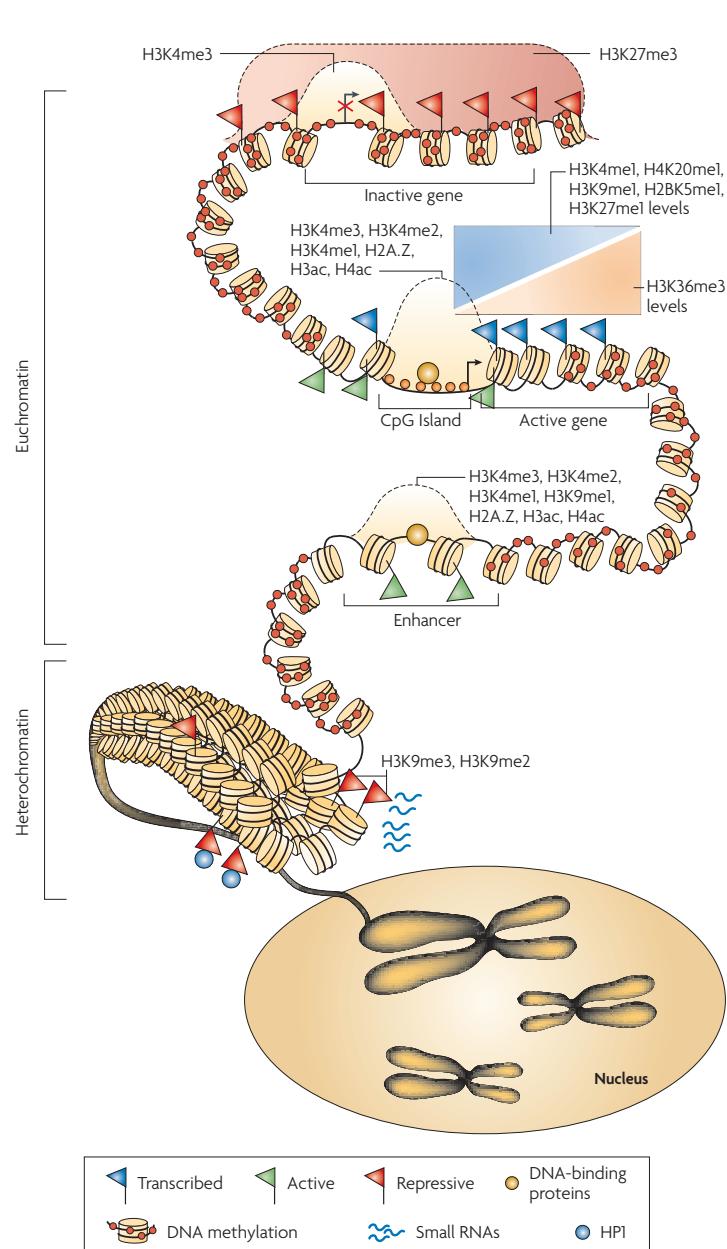


FIGURE 1.6 – Caractéristiques de l'épigénome. Figure tirée de ([Schones and Zhao, 2008](#)). Les chromosomes sont partagés entre régions accessibles d'euchromatine et régions difficilement accessibles d'hétérochromatine. Les régions hétérochromatiques sont marquées par de la di- et triméthylation de la lysine 9 de l'histone H3 (H3K9me2 et H3K9me3). La méthylation de l'ADN est pervasive à travers le génome et est seulement absente dans les régions telles que 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 transcrtes activement à proximité de la région 5' des gènes (en aval), alors que la marque H3K36 marque les gènes transcrits dans leur région 3' (en amont).

lation 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 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 ARNs doubles brins ou dsRNA (*double-stranded RNAs*) qui, une fois clivés par la protéine Dicer, forment des petits peptides de 22 nts appelés siRNAs (*small interfering RNAs*) qui recrutent le complexe protéique RISC (*RNA-induced silencing complex*) et ciblent spécifiquement des ARNm (Hammond et al., 2001; Hannon, 2002). Cette méthode est connue sous le nom d'interférence ARN (RNAi) et est aujourd'hui couramment utilisée pour inhiber l'expression d'un gène. De manière similaire, les microARNs ou miRNAs sont des ARNs de ~ 23 nts issus d'ARNs plus longs appelés « épingle à cheveux » ou *hairpins* qui s'associent à la protéine Argonaute du complexe RISC pour entraîner la dégradation spécifique d'ARNms (Bartel, 2009).

- **Régulation post-traductionnelle**

Les modifications post-traductionnelles affectent les protéines issues de la traduction des ARNs. Ces modifications 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. Par ailleurs, 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).

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. Celle-ci est en effet loin d'être due au hasard. Ainsi, 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

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

Chapitre 1. Introduction générale.

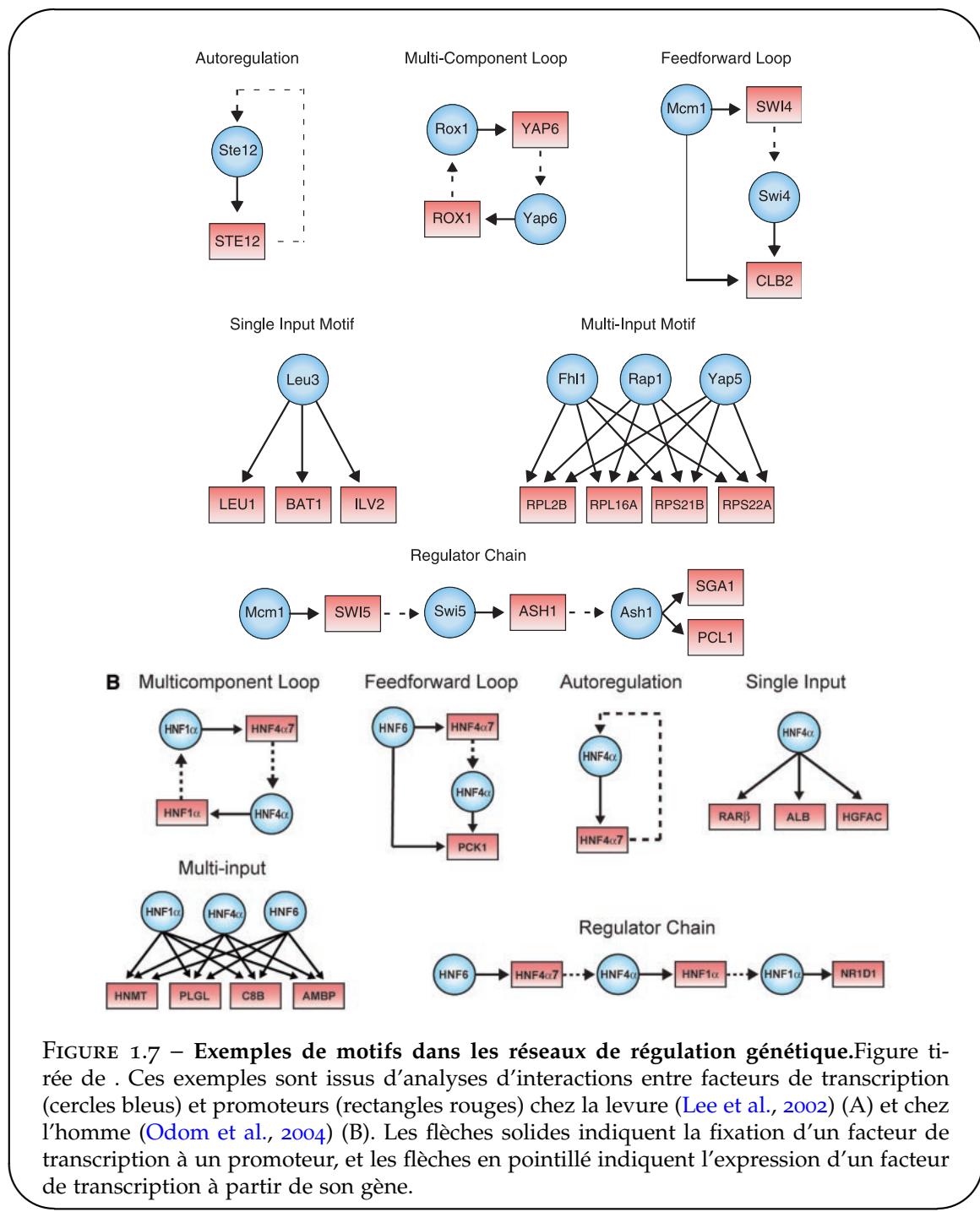


FIGURE 1.7 – Exemples de motifs dans les réseaux de régulation génétique. Figure tirée de . Ces exemples sont issus d'analyses d'interactions entre facteurs de transcription (cercles bleus) et promoteurs (rectangles rouges) chez la levure (Lee et al., 2002) (A) et chez l'homme (Odom et al., 2004) (B). 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.

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

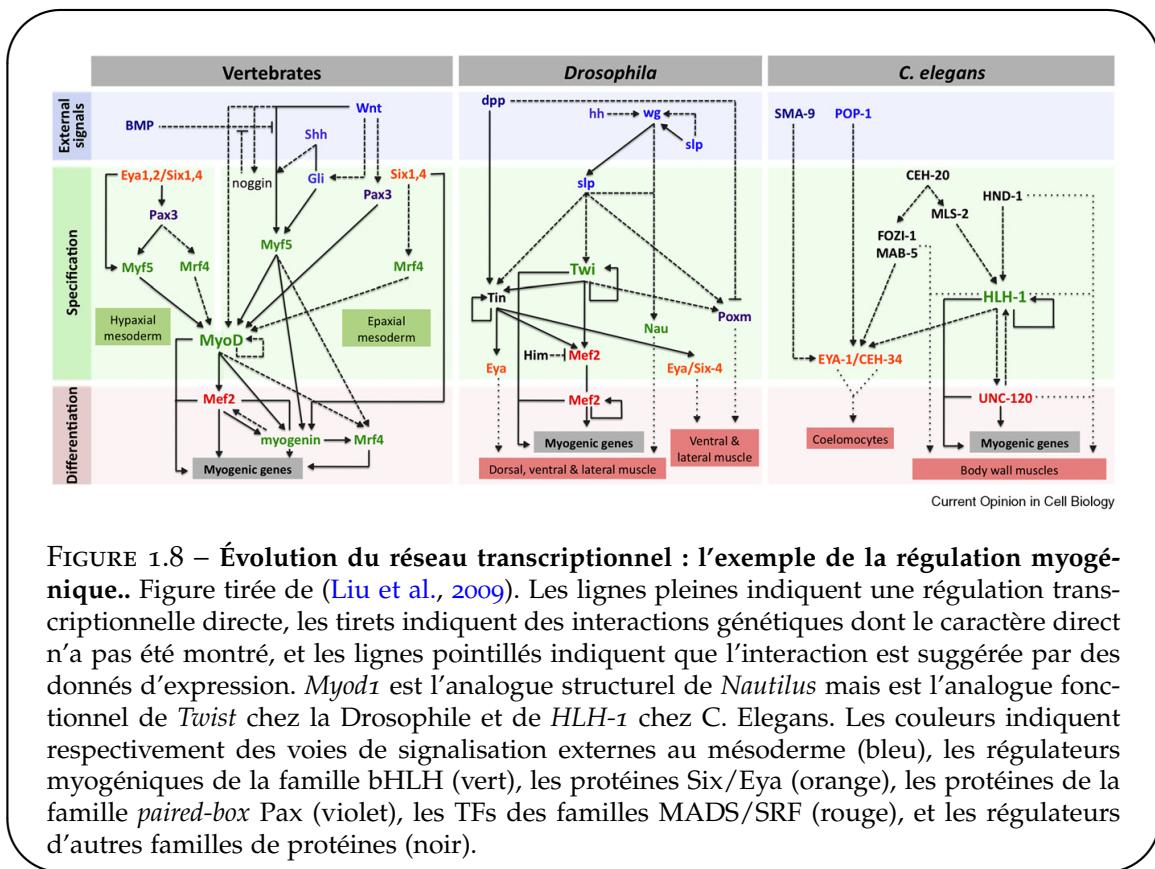


FIGURE 1.8 – Évolution du réseau transcriptionnel : l'exemple de la régulation myogénique.. Figure tirée de (Liu et al., 2009). Les lignes pleines indiquent une régulation transcriptionnelle directe, les tirets indiquent des interactions génétiques dont le caractère direct n'a pas été montré, et les lignes pointillés indiquent que l'interaction est suggérée par des données d'expression. *Myod1* est l'anologue structurel de *Nautilus* mais est l'anologue fonctionnel de *Twist* chez la Drosophile et de *HLH-1* chez *C. Elegans*. Les couleurs indiquent respectivement des voies de signalisation externes au mésoderme (bleu), les régulateurs myogéniques de la famille bHLH (vert), les protéines Six/Eya (orange), les protéines de la famille paired-box Pax (violet), les TFs des familles MADS/SRF (rouge), et les régulateurs d'autres familles de protéines (noir).

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. Par exemple, la modification d'un régulateur, par exemple une mutation d'un certain acide aminé d'un facteur de transcription, aura des conséquences sur l'ensemble des éléments régulés par ce facteur de transcription. 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. Par ailleurs, certains motifs du réseau, comme les boucles d'autorégulation ou les boucles feedforward, peuvent avoir une grande importance fonctionnelle, favorisant leur conservation.

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 coeur 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

Chapitre 1. Introduction générale.

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*, *hlh-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. Ainsi le cœur du réseau est 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), recâblage 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 Modèles mathématiques des interactions protéine-ADN

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.

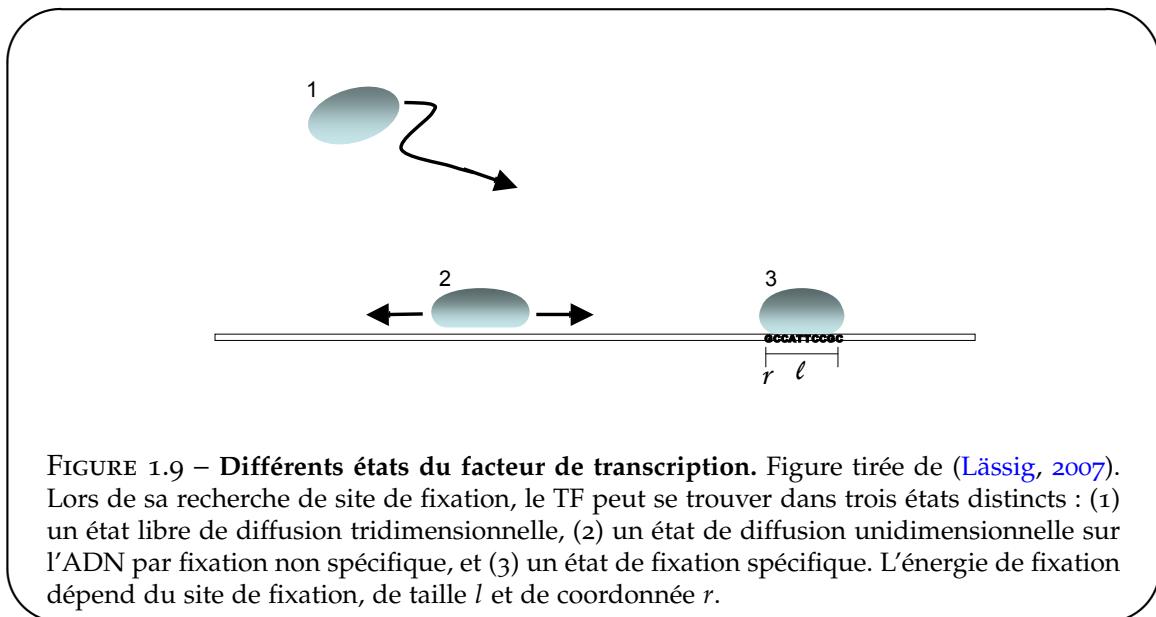


FIGURE 1.9 – Différents états du facteur de transcription. Figure tirée de ([Lässig, 2007](#)). Lors de sa recherche de site de fixation, le TF peut se trouver dans trois états distincts : (1) un état libre de diffusion tridimensionnelle, (2) un état de diffusion unidimensionnelle sur l'ADN par fixation 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 r l’ADN, et une interaction spécifique $E_s(r)$ qui dépend de la séquence de taille $l \sim 10$ à la position r . 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. Au final, le facteur de transcription peut être 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 fonctionnelle ([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 tant 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 les barrières de haute énergie en passant en mode de recherche non spécifique. Ce processus s’avère au final très efficace ([Gerland et al., 2002](#); [Slutsky and Mirny, 2004](#)). 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. 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 TF 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. 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, qu'elle 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 :

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

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 $k_B T$), 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](#)) :

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

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

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

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 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 (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 score, la PWM. Cette matrice peut alors être utilisée pour attribuer un score aux séquences 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

Dans le paragraphe précédent, nous avons vu que 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 $\text{TF} : S_i$ 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 :

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

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

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

où $E_i = -\ln(K_i)$ est l'énergie libre standard de fixation (souvent notée ΔG), et $\mu = \ln[\text{TF}]$ est le potentiel chimique, ces deux quantités étant exprimées en 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 description thermodynamique). 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.

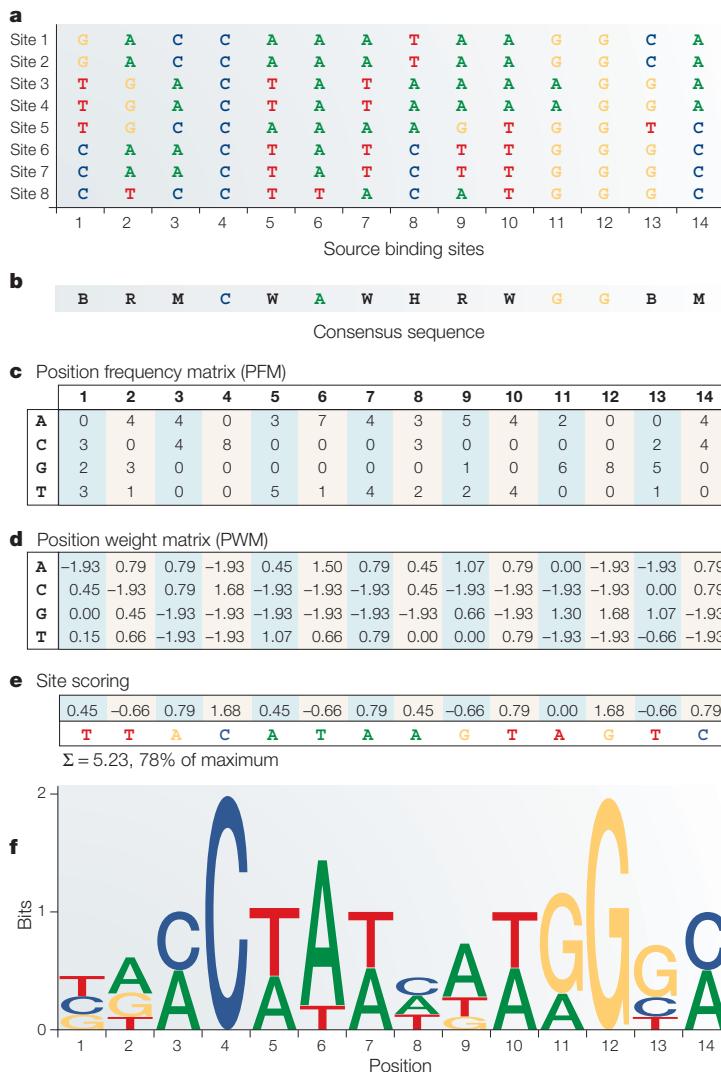


FIGURE 1.10 – Construction et utilisation du modèle PWM. Figure tirée de ([Wasserman and Sandelin, 2004](#)). (a) Supposons connus un certain nombre de sites de fixation d'un facteur de transcription (dans ce cas MEF2). (b) Séquence consensus correspondante utilisant les symboles IUPAC. (c) Une matrice 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 PWMs correspondant. (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 observée.

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. 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 :

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

où les énergies spécifique $E_s(r)$ et non spécifique E_{ns} sont exprimées en unités de $k_B T$. Notons i la position du site fonctionnel. On peut écrire :

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

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

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

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

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

où $F_0 = -\log \mathcal{Z}_0$ 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, et considérons que le nombre de TFs est grand devant le nombre de sites de reconnaissance : 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,

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

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

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

Chapitre 1. Introduction générale.

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 finalement

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

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 bien approximée par (Gerland et al., 2002)

$$\mu = F_0 + \log n \quad (1.15)$$

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 seuil de la fonction de Fermi un facteur $\log n$ par rapport au cas d'un seul TF. Par ailleurs, cette approche thermodynamique nous a permis de généraliser le modèle biophysique simple introduit au paragraphe §1.3.3.

1.4 Mesures expérimentales des interactions protéine-ADN

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, Maerkl 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 (Maerkl 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 ainsi finement contrôlé. Dans ce cas, le fluide contient des gènes synthétique codant pour le TF ainsi que du matériel permettant la synthèse de la protéine au sein de la chambre, évitant l'étape de purification du TF. Chaque chambre du système contient des anticorps attachés 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

1.4. Mesures expérimentales des interactions protéine-ADN

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. Ainsi, un modèle plus raffiné prenant en compte l'énergie d'interaction non spécifique et incluant des interactions entre nucléotides voisins permet de rendre compte des valeurs observées ([Stormo and Zhao, 2007](#)). Nous reviendrons amplement sur la nécessité de prendre en compte les interactions entre paires de nucléotides 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 (SPR en anglais) 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 attachées de l'autre côté de sa surface. Si de l'ADN est attaché à la surface, la fixation du TF induit un changement 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 fixés des brins d'ADN. Une puce contient tous les sites de fixation de 10bp possibles, groupés dans les puits en fonction de leur coeur de 8bp (32,768 séquences en comptant les deux brins) avec possibilité de distinguer les bases flanquantes dans certains cas spécifiés. 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 incluant toutes les séquences de 10 bases possibles : 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. 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](#)).

Chapitre 1. Introduction générale.

- **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 bouts 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 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 avant d'être 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 méthode SELEX a été mise au point, des avancées considérables ont été faites 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 HT-SELEX (Nagaraj et al., 2008), aussi appelée Bind-n-Seq (Zykovitch 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 (~ 10¹⁵ séquences) on peut couvrir l'ensemble des sites de 25bp possibles. Le séquençage haut-débit permet d'en échantillonner ~ 10⁸, ce qui est largement suffisant pour contraindre des modèles d'énergie indépendants, même pour des TFs ayant des sites de fixations de taille > 15bp 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 utilisant 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 récemment pu utiliser 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'intégra-

tion 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 *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 dans lesquelles le facteur de transcription se fixe à la proie expriment *HIS3*, croissent et forment des colonies, d'où 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.

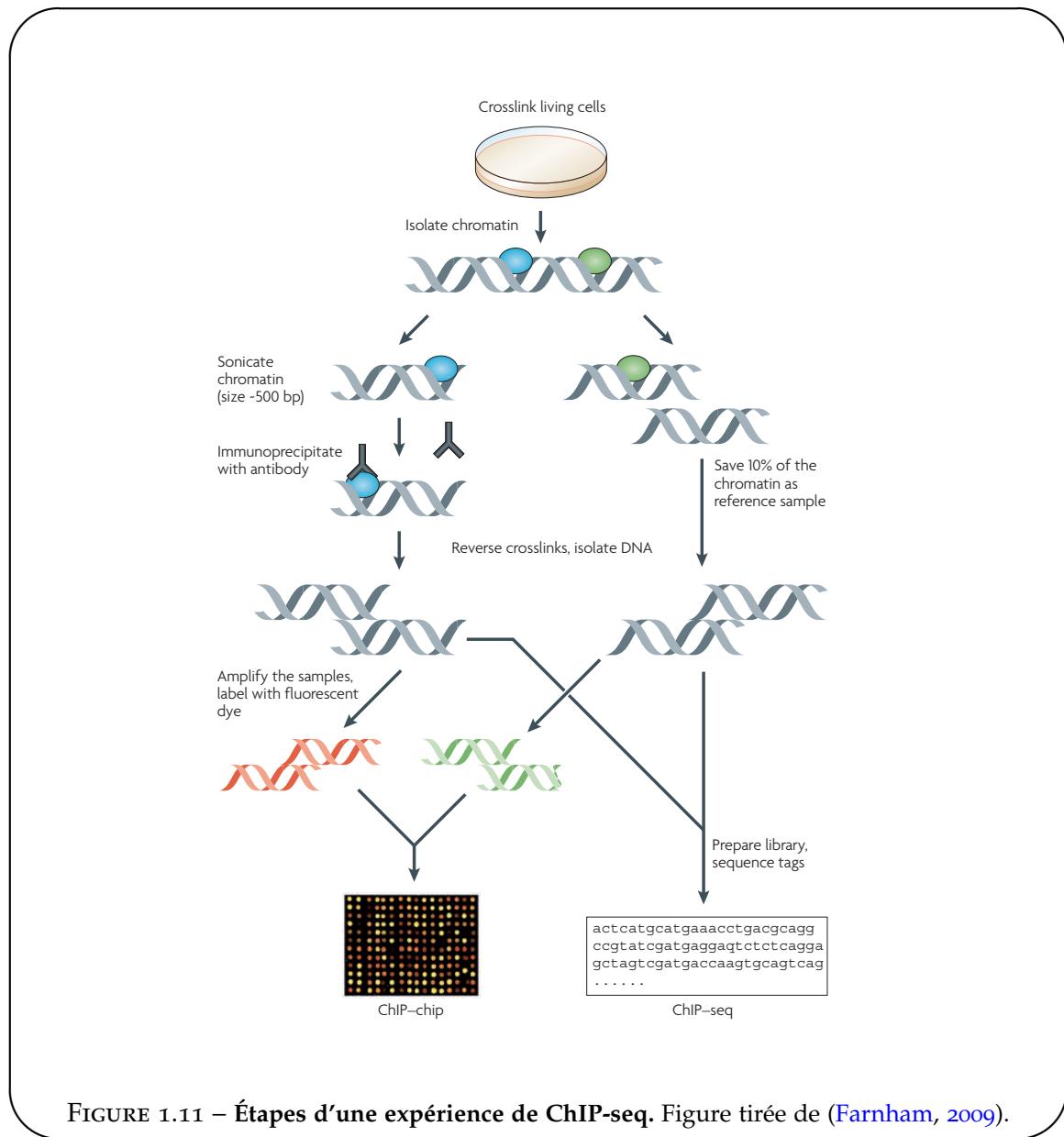
Notons qu'il est aussi possible d'adopter la démarche inverse, c'est-à-dire de partir de quelques sites de fixation présumés fonctionnels mais pour lesquels on ne connaît pas le TF associé. En utilisant une bibliothèque de plasmides codant pour différents TFs hybrides, il est alors possible de déterminer 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 données 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 avec l'ADN en traitant les cellules avec de la formaldéhyde. Cette étape permet de transformer les liaisons faibles 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

FIGURE 1.11 – Étapes d'une expérience de ChIP-seq. Figure tirée de ([Farnham, 2009](#)).

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) ou par séquençage haut débit (ChIP-seq).

Dans le cas de la ChIP-on-chip, l'échantillon immunoprécipité et l'ADN de départ (*input*) sont marqués avec des colorants fluorescents et hybrideront à une puce à ADN composée de très nombreux trous 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'immunoprécipitation et d'*input*.

Dans le cas du ChIP-seq, l'échantillon immunoprécipité est analysé par séquençage à haut

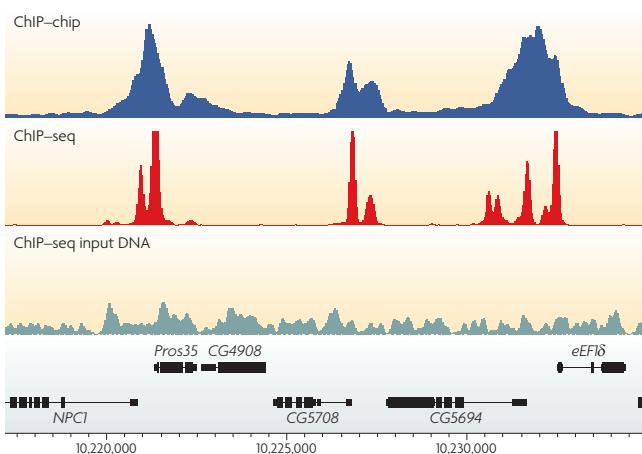


FIGURE 1.12 – Résolution des expériences ChIP-on-chip et ChIP-seq. Figure tirée de (Park, 2009). Exemples de profils de fixation de la protéine à chromodomaine Chromator générés à partir d'expériences de ChIP-on-chip (intensité relative par rapport au contrôle, bleu) et de ChIP-seq (densité de séquences, 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.

débit, résultant en une librairie de *reads* d'une longueur typique variant entre 27 et 50bp issus des bouts 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 ($\leq 100\text{nt}$) 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 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'avoir un ensemble de 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

Chapitre 1. Introduction générale.

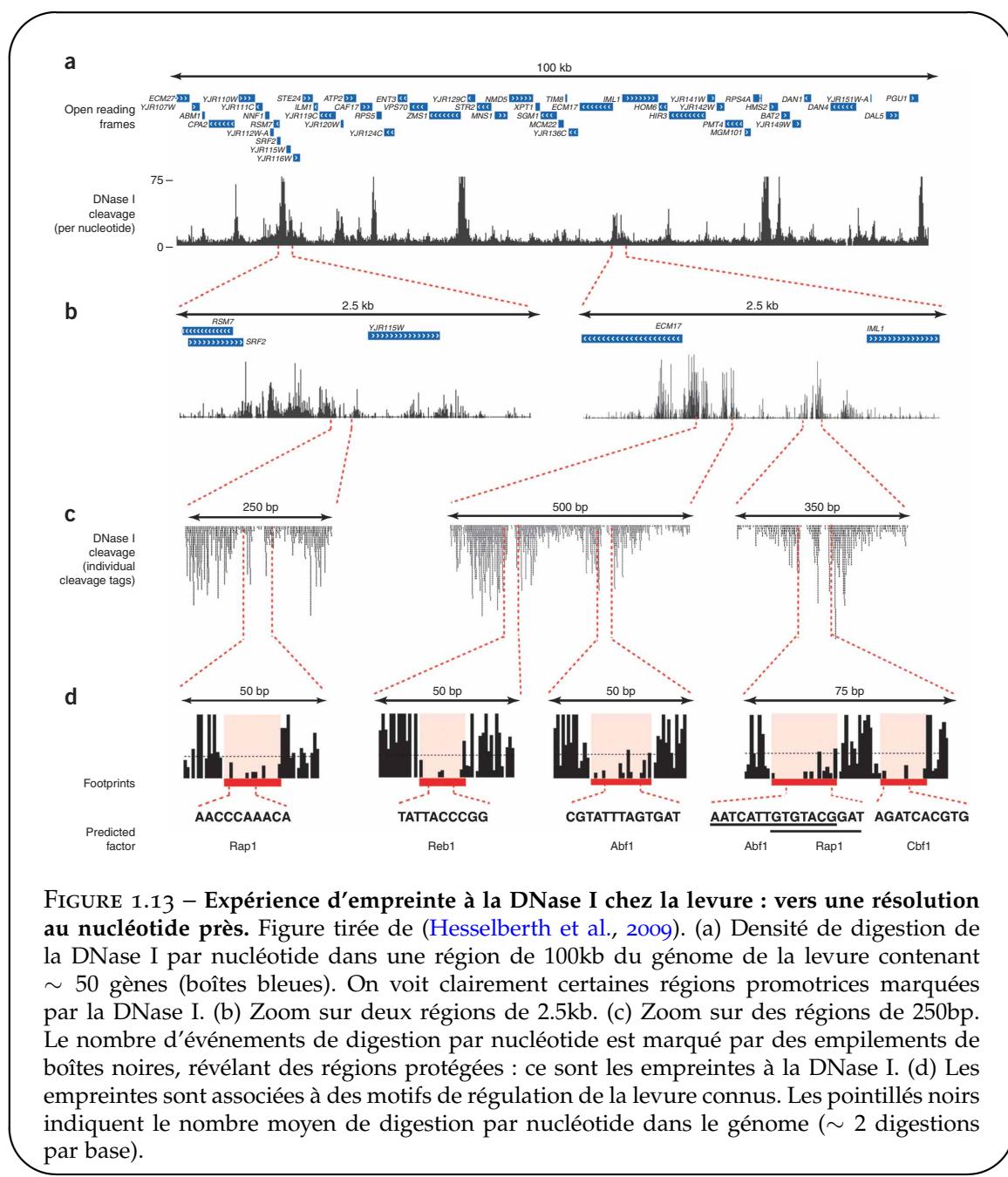


FIGURE 1.13 – Expérience d’empreinte à la DNase I chez la levure : vers une résolution au nucléotide près. Figure tirée de ([Hesselberth et al., 2009](#)). (a) Densité de digestion de la DNase I par nucléotide dans une région de 100kb du génome de la levure contenant ~ 50 gènes (boîtes bleues). On voit 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).

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 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 (~ 15bp) 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

Nous l'avons vu en 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 être au cœur de nombreuses pathologies, liées pour la plupart à une expression génétique aberrante. Par exemple, 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 proposons un survol des différents types de CRMs, de leur structure, leur évolution et leur prédiction.

1.5.1 Les différents types de CRMs

La régulation de l'expression génétique implique l'interaction entre des facteurs de transcription et des CRMs. Selon leur rôle dans la régulation de l'expression génétique, les CRMs peuvent être distingués en trois catégories.

- **Promoteurs**

Les promoteurs permettent la fixation de l'ARN polymérase pour débuter la formation d'un transcrit 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 ~ 100bp 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 sur 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 possèdent 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 1.5.7). 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

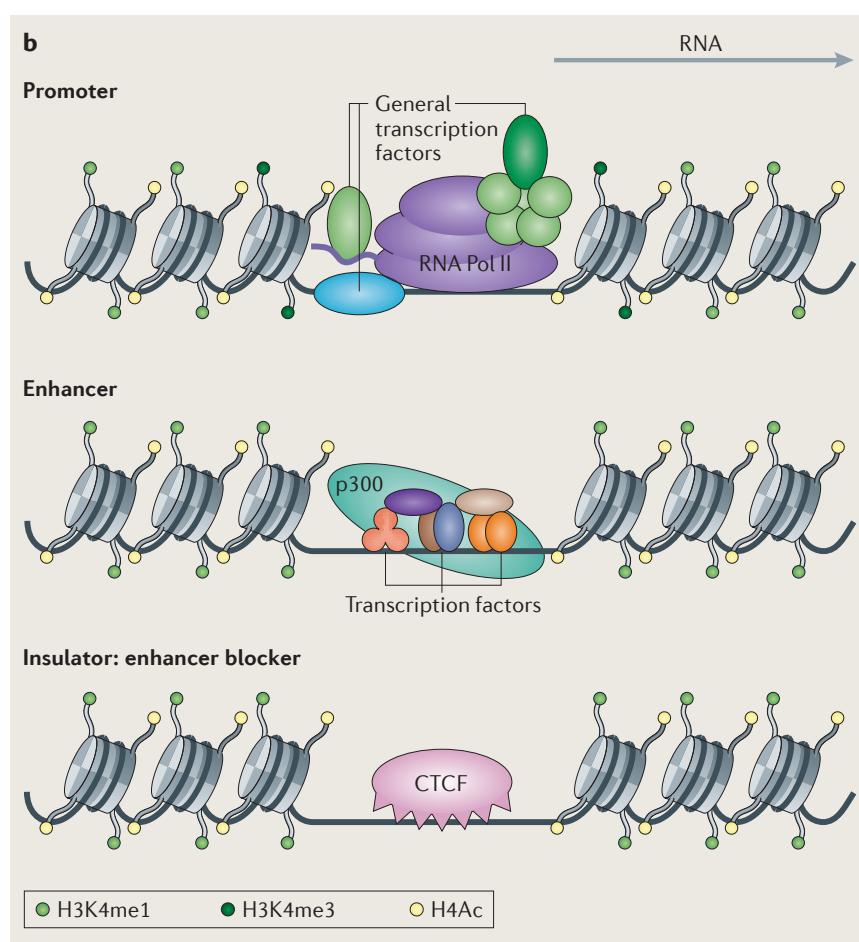


FIGURE 1.14 – Les différents types de CRMs et leurs marques épigénétiques.

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 sont marqués 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.

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). 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 à différents tissus, comme cela l'a été montré dans le cas du cluster des gènes de détermination myogénique *Myf5* et *Mrf4* (Carvajal et al., 2008). 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'enhaner, 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 aussi servir de barrière de protection contre des marques d'hétérochromatine répressives. De tels insulateurs permettent de manière pratique 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. Cette activité de barrière de protection peut très bien s'associer à celle de blocage d'enhaner.

De même que les enhancers, les insulateurs peuvent se situer à des distances variables des gènes qu'ils régulent. Par ailleurs, bien que la fixation de CTCF soit requise chez les mammifères, cette protéine possède d'autres fonctions que celle d'isolation, et tous les sites de CTCF ne correspondent pas forcément à des insulateurs (Phillips and Corces, 2009).

1.5.2 Encodage de patterns spatiaux

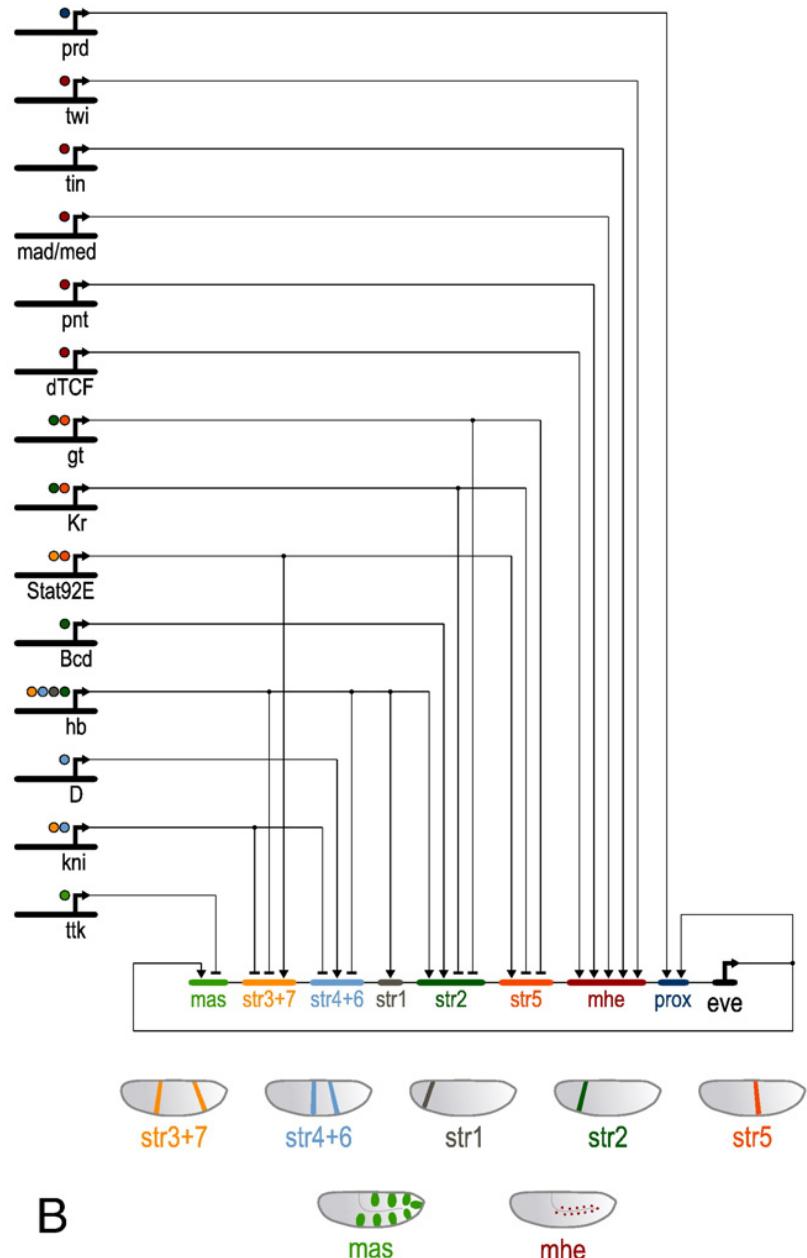
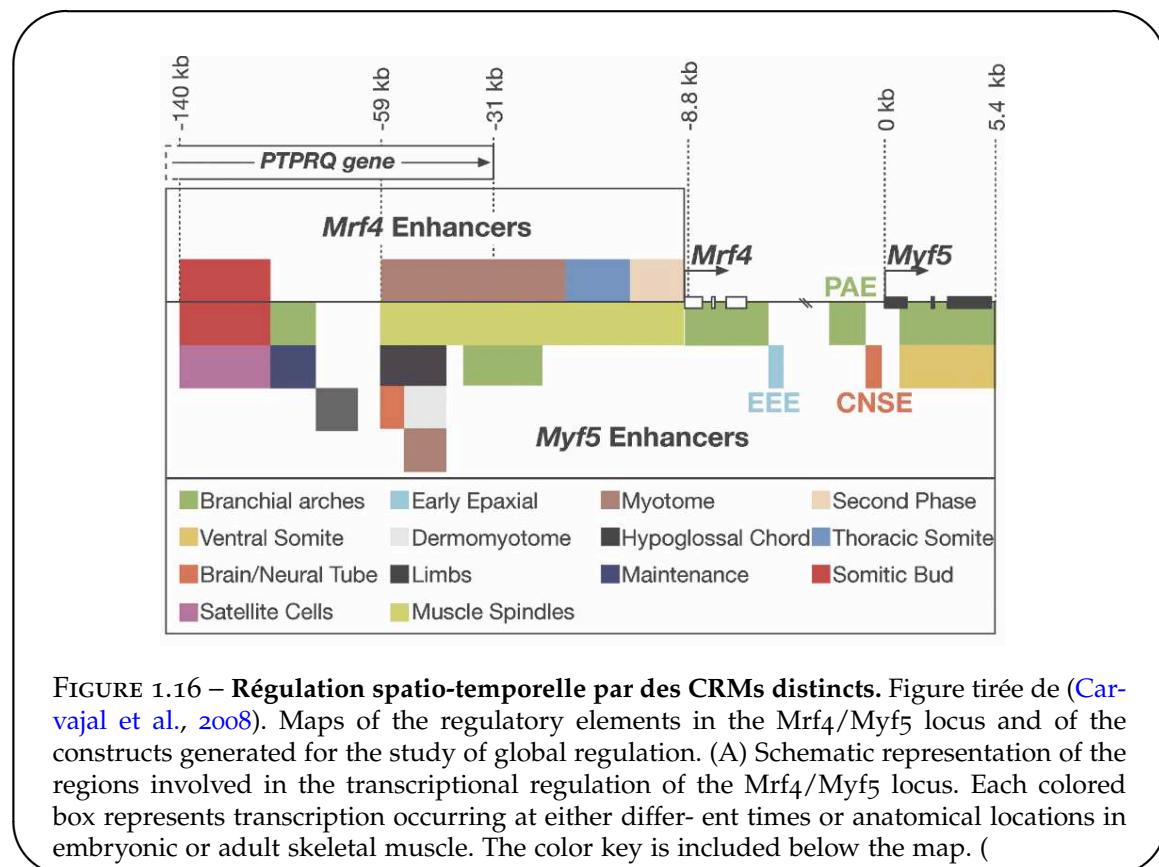


FIGURE 1.15 – Différents CRMs conduisent à différents patterns d’expression.

Figure tirée de ([Wilczynski and Furlong, 2010](#)), montrant le réseau de cis-régulation du gène *even-skipped* chez la *Drosophila*. Des enhancers individuels (boîtes colorées) conduisent à des motifs d’expression distincts (indiqués dans les embryons par la même couleur). Les cercles colorés au sein des régulateurs indiquent les couleurs des CRMs régulés.



1.5.3 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. 1.17).

- **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 de sites de liaisons pour différents TFs.

- **Le modèle “enhanceosome”**

Dans ce modèle, l'architecture des sites de liaison est de prime importance. Le paradigme en est l'enhancer du gène humain interferon- β , sur lequel 8 TFs se fixent pour former une 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. 1.18).

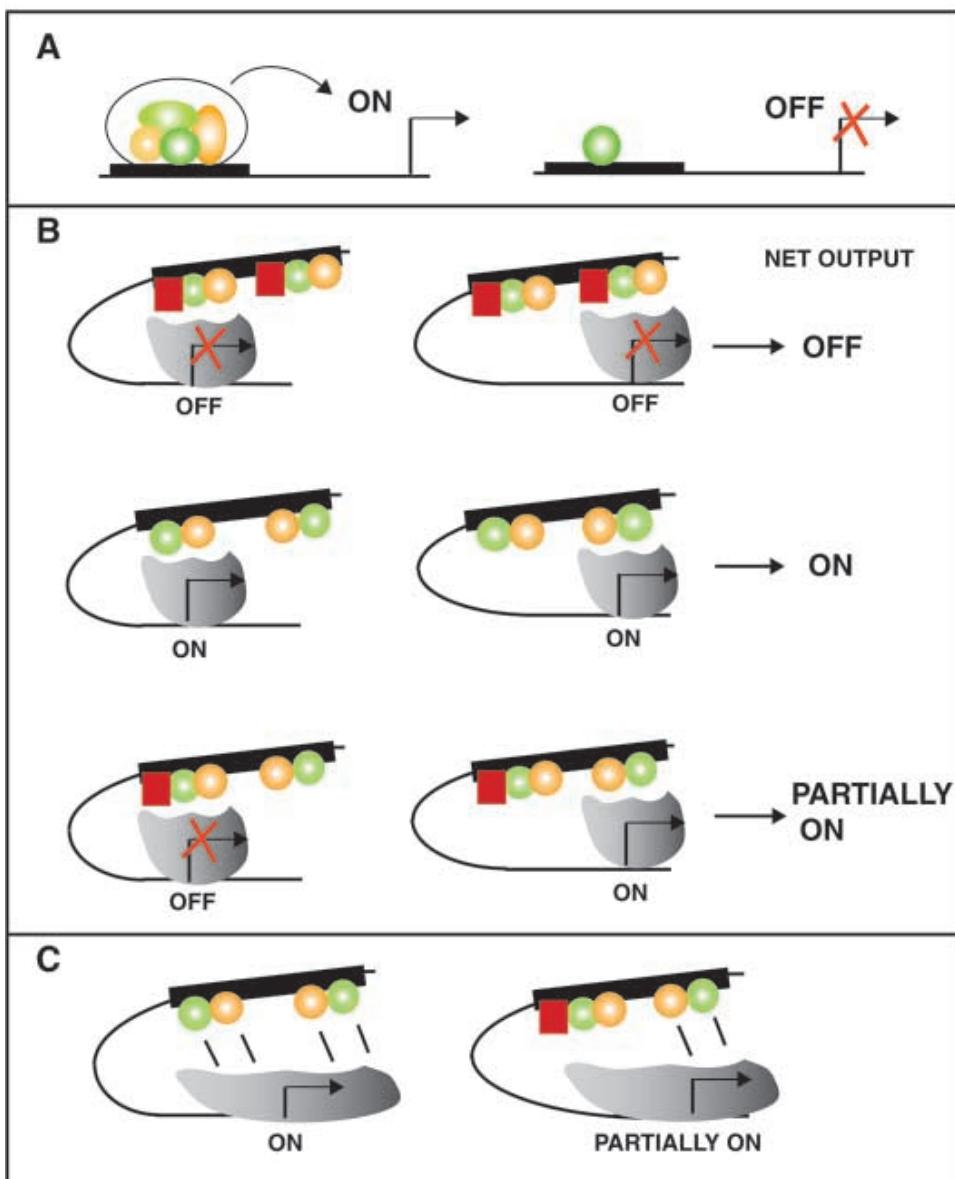


FIGURE 1.17 – Deux modèles d’enhancers : enhanceosome et billboard.

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 ou enhanceosome 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 des sous-unités peuvent contenir différentes informations (répression ou activation par exemple) que la machinerie basale échantillonne soit itérativement (B), soit simultanément (C).

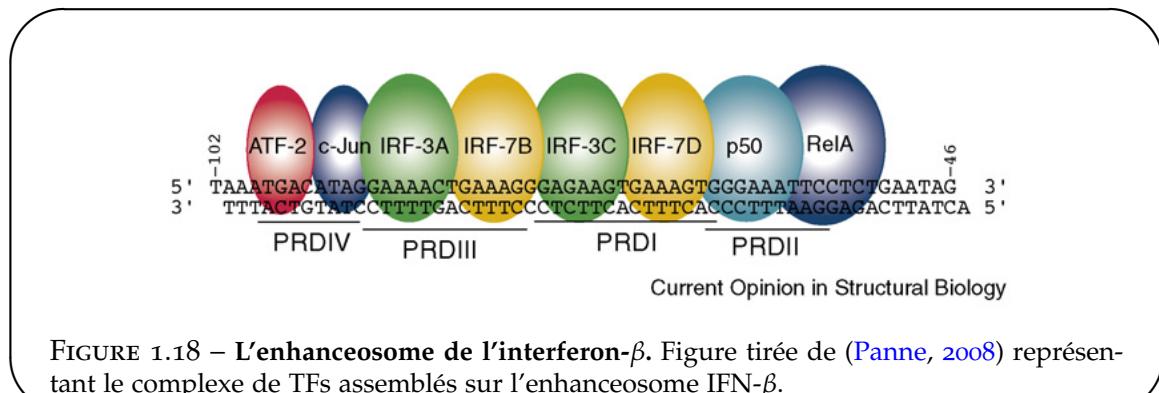


FIGURE 1.18 – L’enhanceosome de l’interferon- β . Figure tirée de (Panne, 2008) représentant le complexe de TFs assemblés sur l’enhanceosome IFN- β .

1.5.4 Évolution des enhancers

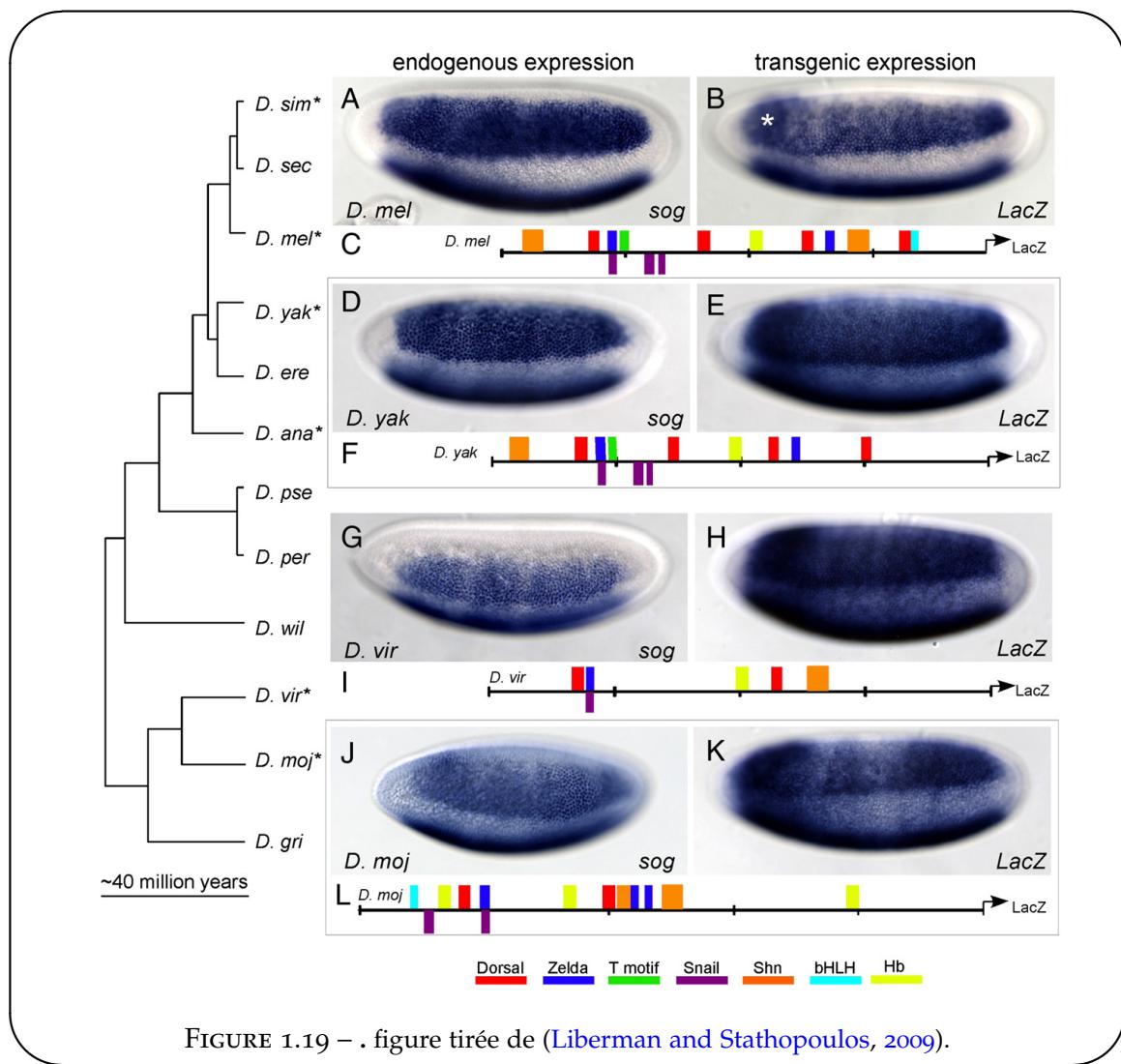


FIGURE 1.19 – . figure tirée de (Liberman and Stathopoulos, 2009).

(Feschotte, 2008)

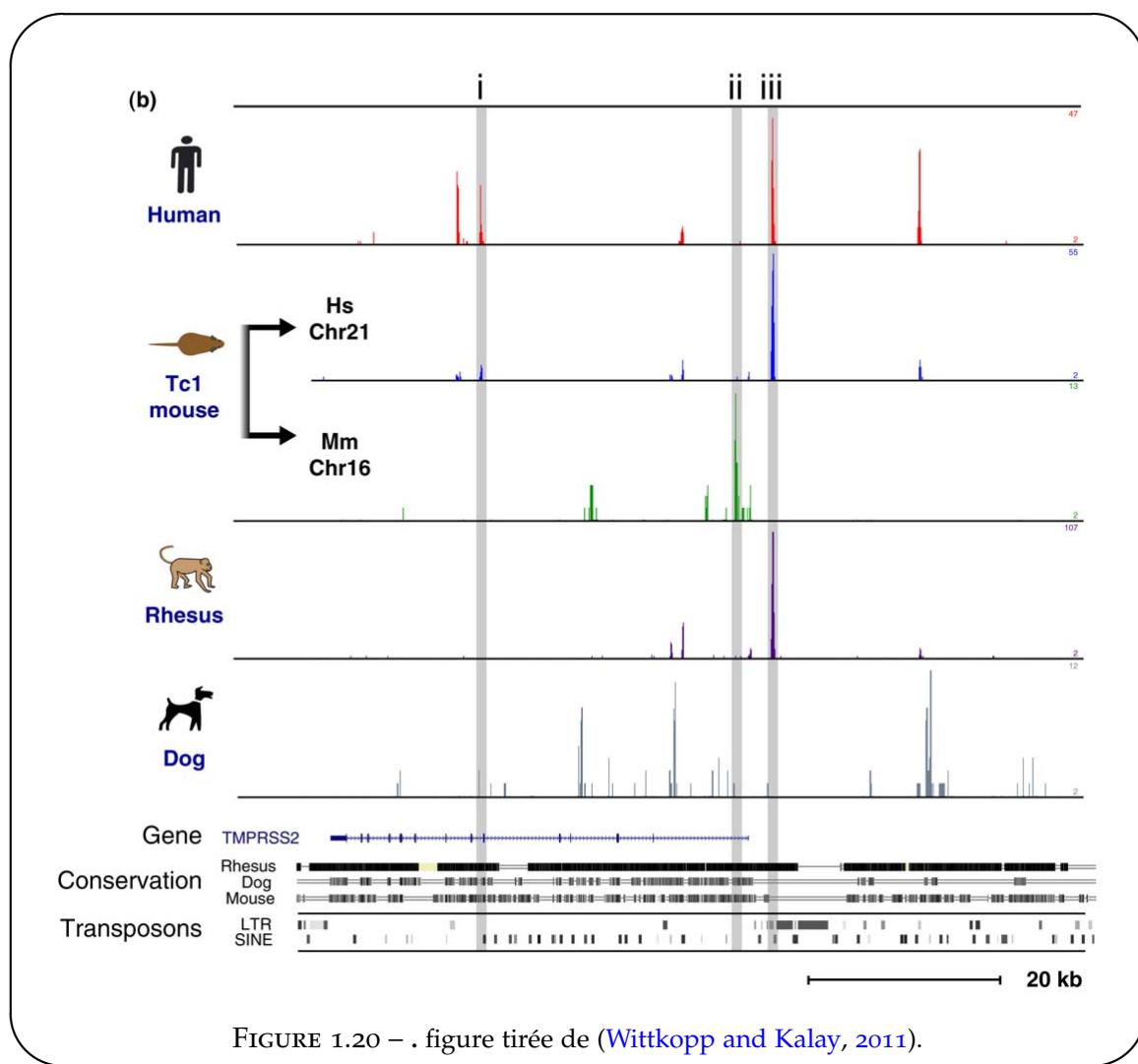


FIGURE 1.20 – . figure tirée de (Wittkopp and Kalay, 2011).

1.5.5 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. Récemment, 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 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 gap de la Drosophile possèdent tous des « shadow enhancers », dont le rôle semble être d'assurer une plus grande précision spatiale du motif d'expression du gène régulé ([Perry et al., 2011](#)). La perte de l'un des CRMs, proximal comme « shadow enhancers », conduit à une expression trop restreinte ou trop répandue spatialement selon le cas.

1.5.6 Les super enhancers

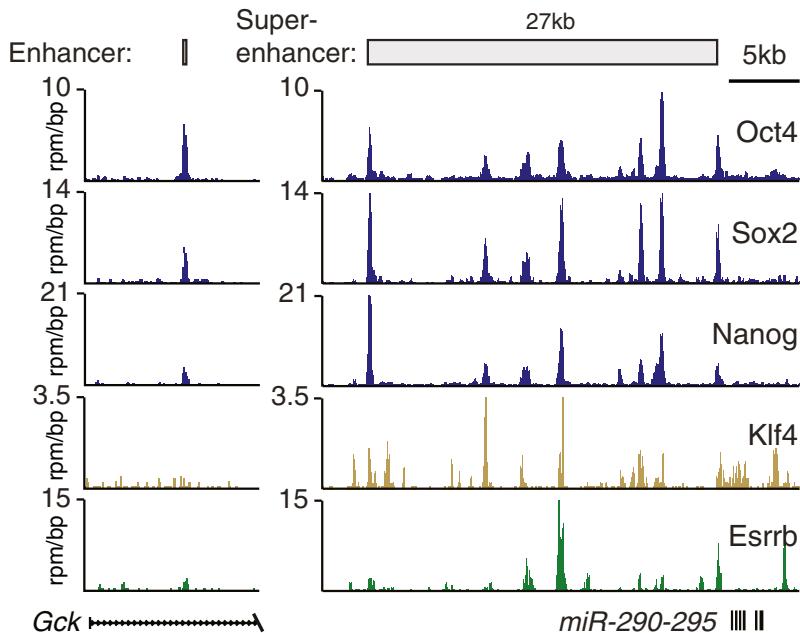


FIGURE 1.21 – De l’enhancer au super-enhancer.

Figure tirée de ([Whyte et al., 2013](#)), montrant les profils de ChIP-seq de Oct4, Sox2, Nanog, Klf4 et Esrrb aux loci de *Gck* et *miR-290-295* dans les cellules souches embryonnaires.

1.5.7 Validation expérimentale

Lorsqu’un CRM est prédict, il existe plusieurs méthodes pour s’assurer de sa fonctionnalité.

Tout d’abord, une méthode indirecte donnant du crédit à la prédition 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 prédir le CRM (par exemple, s’exprimer dans le muscle pour une prédition de CRMs utilisant l’abondance de sites de liaison de TFs musculaires), 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. 1.22). 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.

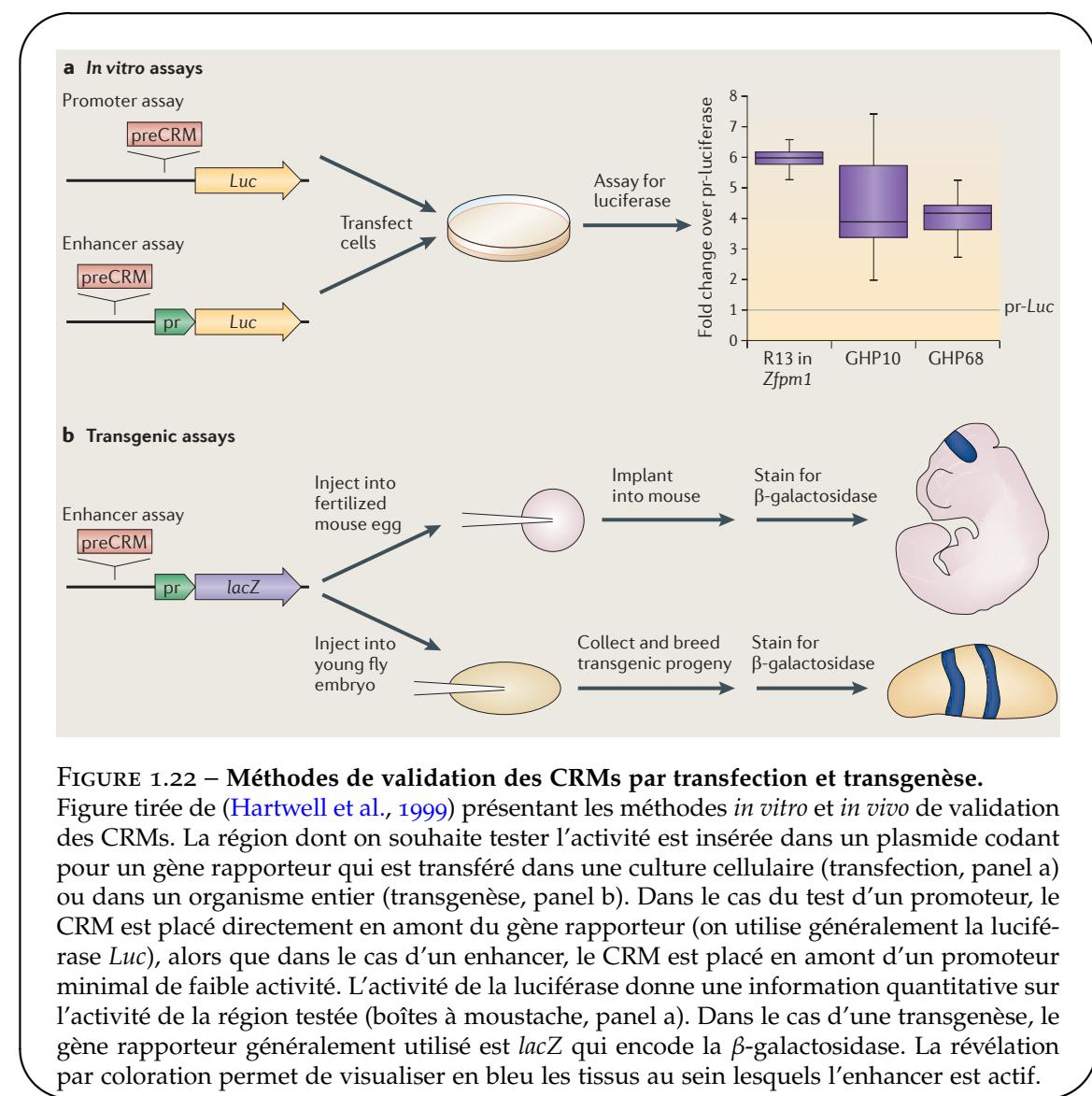


FIGURE 1.22 – Méthodes de validation des CRMs par transfection et transgenèse.
 Figure tirée de (Hartwell et al., 1999) présentant les méthodes *in vitro* et *in vivo* de validation des CRMs. La région dont on souhaite tester l'activité est insérée dans un plasmide codant pour un gène rapporteur qui 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 desquels l'enhancer est actif.

1.5.8 Prédiction des CRMs

1.6 Banques de données

1.6.1 Séquences génomiques et alignements

statistiques du genome (lognormal)

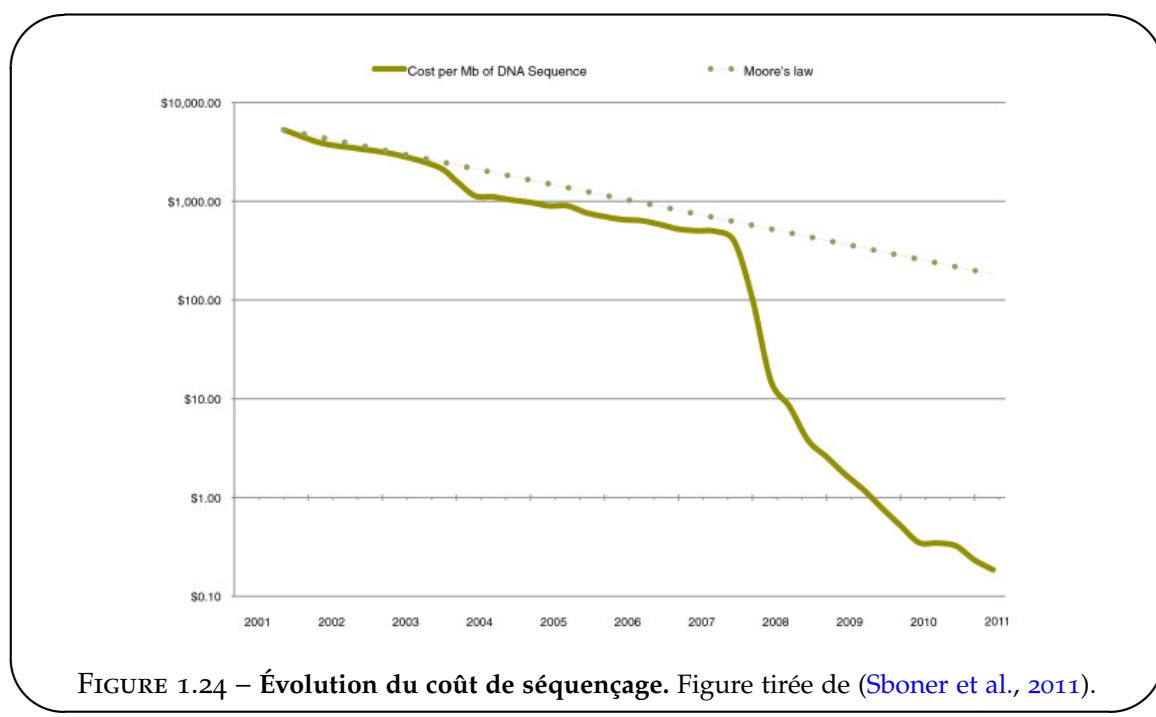
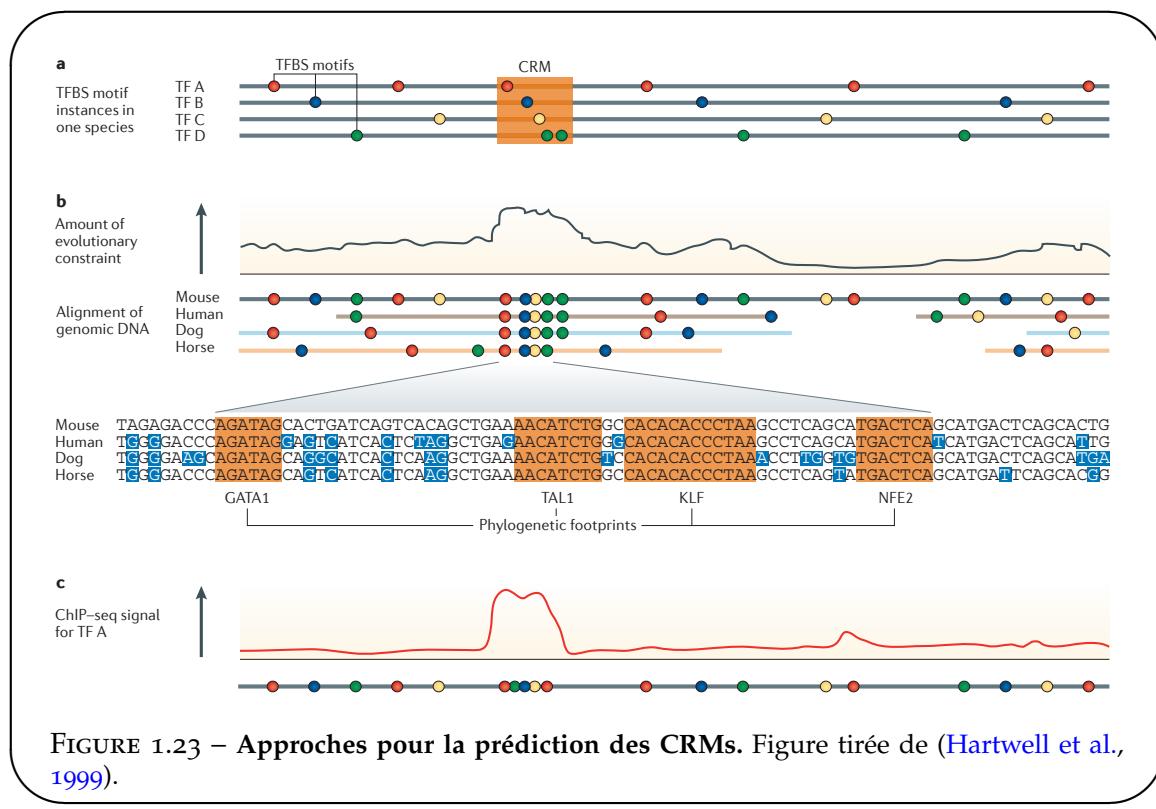
1.6.2 Annotations (TSSs, repeats...)

1.6.3 Jaspar et Transfac

1.6.4 Visualisation sur UCSC

1.6.5 Le projet ENCODE

Chapitre 1. Introduction générale.



Chapitre 2

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

3.1 47

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 (Consortium, 2011) accessibles à partir du site internet de UCSC³, 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.

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., 2010).

2.2.2 Statistique « background » des séquences

Présence de corrélations.

2.3 Présentation de l'algorithme

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*

³. <http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeCaltechTfbs/>

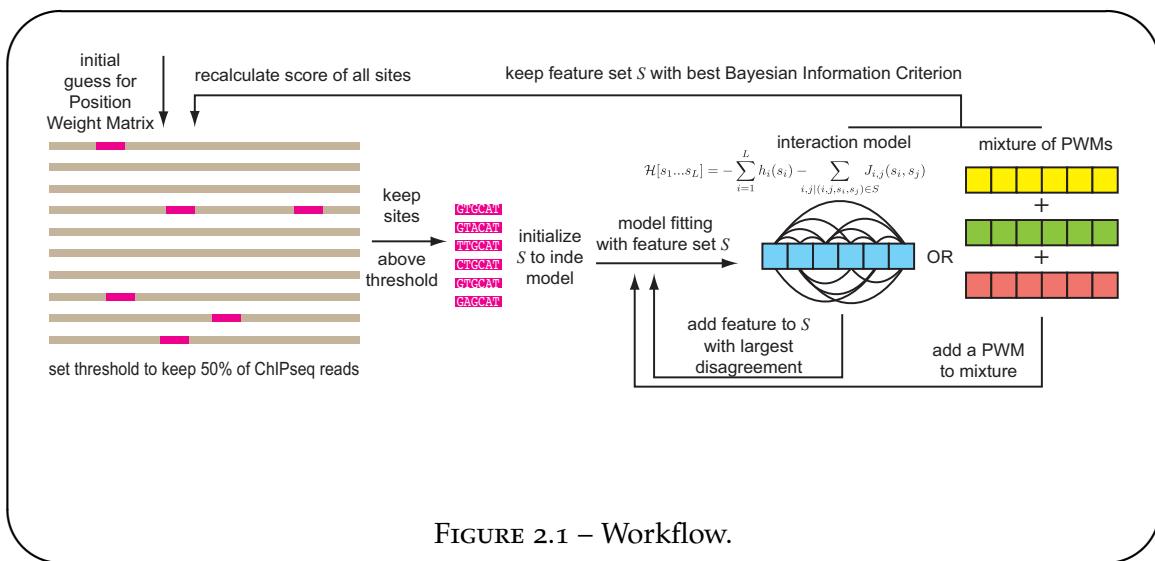
2.6. Comparaison avec des données *in vitro*

FIGURE 2.1 – Workflow.

2.6.1 Conclusion de la section 2.6

thèse : version du lundi 17 juin 2013 à 17 h 53

Chapitre 2. Modèles de fixation des Facteurs de Transcription à l'ADN.

Chapitre 3

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

5.1	55
-----	-------	----

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

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 31.)
- Alon, U. (2007a). An introduction to systems biology : Design principles of biological circuits (mathematical and computational biology series vol 10). (Page 11.)
- Alon, U. (2007b). Network motifs : theory and experimental approaches. *Nat Rev Genet*, 8(6) :450–461. (Page 12.)
- Bartel, D. P. (2009). Micrornas : target recognition and regulatory functions. *Cell*, 136(2) :215–33. (Page 11.)
- Baylies, M. K., Bate, M., and Ruiz Gomez, M. (1998). Myogenesis : a view from drosophila. *Cell*, 93(6) :921–7. (Page 13.)
- 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 29.)
- Berg, O. and von Hippel, P. (1987). Selection of dna binding sites by regulatory proteins : Statistical-mechanical theory and application to operators and promoters. *Journal of molecular biology*, 193(4) :723–743. (Page 16.)
- 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 15.)
- Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, 3rd, P. W., and Bulyk, M. L. (2006). Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11) :1429–35. (Page 21.)
- Bird, A. (2002). Dna methylation patterns and epigenetic memory. *Genes Dev*, 16(1) :6–21. (Page 11.)
- 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 22.)
- 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 6.)
- 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 4.)
- 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 21.)

Bibliographie

- 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 21.)
- Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engström, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A., and Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6) :626–35. (Page 27.)
- 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 28 et 31.)
- Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N. D., Wei, C.-L., and Ng, H.-H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6) :1106–17. (Page 42.)
- 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 29.)
- Consortium, E. P. (2011). A user's guide to the encyclopedia of dna elements (encode). *Plos Biol*, 9(4) :e1001046. (Page 42.)
- 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 7.)
- 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 17.)
- Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. *Nat Rev Genet*, 10(9) :605–16. (Page 24.)
- Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*, 9(5) :397–405. (Page 33.)
- 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 22.)
- 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 35.)
- Furusawa, C. and Kaneko, K. (2012). A dynamical-systems view of stem cell biology. *Science*, 338(6104) :215–217. (Page 5.)

- 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 15, 17, 18, 19 et 20.)
- Giocomo, L. M., Moser, M.-B., and Moser, E. I. (2011). Computational models of grid cells. *Neuron*, 71(4) :589–603. (Page 18.)
- Graf, T. and Enver, T. (2009). Forcing cells to change lineages. *Nature*, 462(7273) :587–94. (Page 7.)
- 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 11.)
- Gurdon, J. B. and Melton, D. A. (2008). Nuclear reprogramming in cells. *Science*, 322(5909) :1811–5. (Page 7.)
- 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 11.)
- Hannon, G. J. (2002). Rna interference. *Nature*, 418(6894) :244–51. (Page 11.)
- Hardison, R. C. and Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics*, 13(7) :469–483. (Page 28.)
- Hartwell, L., Hopfield, J., Leibler, S., and Murray, A. (1999). From molecular to modular cell biology. *Nature*, 402(6761) :47. (Pages 37 et 38.)
- Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243) :108–12. (Page 29.)
- 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 26.)
- Hong, J.-W., Hendrix, D. A., and Levine, M. S. (2008). Shadow enhancers as a source of evolutionary novelty. *Science*, 321(5894) :1314. (Page 34.)
- Jaynes, E. (1957). Information theory and statistical mechanics. ii. *Physical review*, 108(2) :171. (Page 41.)
- 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 22.)
- Jolma, A., Yan, J., Whitington, 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 22.)

Bibliographie

- Kaufmann, S. (1993). The origins of order. (Page 6.)
- 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 4.)
- 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 21.)
- 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 11.)
- 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 7.)
- Kulkarni, M. M. and Arnosti, D. N. (2003). Information display by transcriptional enhancers. *Development*, 130(26) :6569–75. (Pages 31 et 32.)
- Lässig, M. (2007). From biophysics to evolutionary genetics : statistical aspects of gene regulation. *BMC Bioinformatics*, 8(Suppl 6) :S7. (Pages 15 et 19.)
- 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 11 et 12.)
- 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 the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*, 12(14) :1725–35. (Page 28.)
- Liberman, 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. (Page 33.)
- 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 13.)
- Maerkl, S. and Quake, S. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809) :233. (Page 20.)
- Maniatis, T., Goodbourn, S., and Fischer, J. A. (1987). Regulation of inducible and tissue-specific gene expression. *Science*, 236(4806) :1237–45. (Page 28.)
- 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 35.)
- 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 11.)

- 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. (Page 22.)
- Nurse, P. and Hayles, J. (2011). The cell in an era of systems biology. *Cell*. (Page 8.)
- 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 11 et 12.)
- 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 22.)
- Ondek, B., Gloss, L., and Herr, W. (1988). The sv40 enhancer contains two distinct levels of organization. *Nature*, 333(6168) :40–5. (Page 28.)
- Panne, D. (2008). The enhanceosome. *Curr Opin Struct Biol*, 18(2) :236–42. (Pages 31 et 33.)
- Park, P. J. (2009). Chip-seq : advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10) :669–80. (Page 25.)
- 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 35.)
- Phillips, J. E. and Corces, V. G. (2009). Ctcf : master weaver of the genome. *Cell*, 137(7) :1194–211. (Page 29.)
- 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 29.)
- 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 22.)
- 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 38.)
- 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 28.)
- 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 29.)
- Schones, D. E. and Zhao, K. (2008). Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet*, 9(3) :179–91. (Page 10.)
- Shannon, C. (1948). A mathematical theory of communication. *Bell Syst Tech J*, 27(4) :623–656. (Page 41.)

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 11.)
- 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 21.)
- 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 41.)
- 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 15.)
- Smit, A. F. A., Hubley, R., and Green, P. (1996-2010). Repeatmasker open-3.0. <http://www.repeatmasker.org>. (Page 42.)
- 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 16.)
- Stormo, G. D. and Zhao, Y. (2007). Putting numbers on the network connections. *Bioessays*, 29(8) :717–21. (Page 21.)
- Stormo, G. D. and Zhao, Y. (2010). Determining the specificity of protein-dna interactions. *Nature Reviews Genetics*, 11(11) :751–60. (Pages 20 et 23.)
- 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 7.)
- Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414) :75–82. (Page 26.)
- 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 22.)
- 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 8 et 9.)
- 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 8.)

- 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. (Page 5.)
- Wallace, J. A. and Felsenfeld, G. (2007). We gather together : insulators and genome organization. *Curr Opin Genet Dev*, 17(5) :400-7. (Page 29.)
- 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 29.)
- Warren, C. L., Kratochvil, N. C. S., Hauschild, K. E., Foister, S., Brezinski, M. L., Dervan, P. B., Phillips, Jr, G. N., 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 21.)
- Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4) :276-87. (Page 18.)
- 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 13.)
- 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. (Page 36.)
- 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 30.)
- 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 15.)
- 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 15.)
- Wittkopp, P. J. and Kalay, G. (2011). Cis-regulatory elements : molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1) :59-69. (Page 34.)
- 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 22.)
- 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 34.)
- 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 25.)

Bibliographie

- Zhao, Y., Granas, D., and Stormo, G. D. (2009). Inferring binding energies from selected binding sites. *PLoS Comput Biol*, 5(12) :e1000590. (Page [17](#).)
- 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. (Page [42](#).)
- 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 [22](#).)

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 17 juin 2013 à 17 h 53*