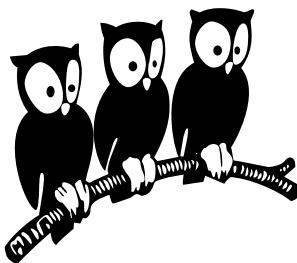


Département de Physique  
École Normale Supérieure

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 des drosophiles et des mammifères**

---

Soutenue le ZZ septembre 2013  
devant le jury composé de :

M.	Emmanuel BARILLOT .....	Examinateur
M.	Vincent HAKIM .....	Directeur de thèse
M.	Pascal MAIRE .....	Examinateur
M.	Massimo VERGASSOLA .....	Rapporteur
M.	Martin WEIGT .....	Rapporteur
M.	Alain ZIDER .....	Examinateur

**thèse**:version du mercredi 17 juillet 2013 à 10 h 31

# Remerciements

---

...

**thèse**:version du mercredi 17 juillet 2013 à 10 h 31

---

# Table des matières

---

<b>Liste des figures</b>	vii
<b>Principales abréviations utilisées</b>	ix
<b>Avant-propos</b>	1
<b>Chapitre 1 - Introduction générale.</b>	3
1.1 Le phénotype cellulaire . . . . .	5
1.2 Les réseaux de régulation génétique . . . . .	10
1.3 Les interactions protéine-ADN : modèles mathématiques . . . . .	20
1.4 Les interactions protéine-ADN : mesures expérimentales . . . . .	27
1.5 Les modules de cis-régulation (CRMs) . . . . .	37
1.6 Prédiction et validation des CRMs . . . . .	50
1.7 Bases de données . . . . .	59
<b>Chapitre 2 - Modèles de fixation des Facteurs de Transcription à l'ADN.</b>	65
2.1 Observations de corrélations au sein des TFBS . . . . .	66
2.2 Modèles existants permettant de décrire la statistique des TFBS . . . . .	67
2.3 Modèles de maximum d'entropie . . . . .	72
2.4 Article . . . . .	77
2.5 Analyse thermodynamique des modèles . . . . .	77
2.6 Conclusion et perspectives . . . . .	80
<b>Chapitre 3 - <i>Imogene</i> : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle</b>	83
3.1 Quelques approches existantes pour la recherche de motifs et de modules de régulation . . . . .	84
3.2 Article . . . . .	96
3.3 Calcul de la moyenne de la postérieure par une méthode MCMC . . . . .	96
3.4 Conclusion et perspectives . . . . .	106
<b>Chapitre 4 - Étude de la différenciation épidermale chez la drosophile</b>	107
4.1 Concept d'optimum de Pareto . . . . .	109
4.2 Article . . . . .	111
4.3 Conclusion et perspectives . . . . .	112

Table des matières

---

<b>Chapitre 5 - Étude de la différenciation musculaire chez la souris</b>	<b>113</b>
5.1 Obtention d'une PWM pour MEF3 . . . . .	115
5.2 Article . . . . .	118
<b>Annexe A - Statistiques génomiques</b>	<b>121</b>
<b>Bibliographie</b>	<b>125</b>

---

# Liste des figures

---

<b>Introduction générale.</b>	<b>3</b>
1.1 Le paysage de la différenciation cellulaire . . . . .	6
1.2 Spécification spatio-temporelle du type cellulaire . . . . .	8
1.3 Différents exemples de reprogrammation cellulaire . . . . .	9
1.4 Vision cybernétique du traitement de l'information par la cellule . . . . .	11
1.5 Un réseau de régulation génétique type . . . . .	12
1.6 Caractéristiques de l'épigénome . . . . .	15
1.7 Exemples de motifs dans les réseaux de régulation génétique . . . . .	16
1.8 Évolution du réseau transcriptionnel : l'exemple de la régulation myogénique. .	18
1.9 Différents états du facteur de transcription . . . . .	20
1.10 Construction et utilisation du modèle PWM . . . . .	23
1.11 Étapes d'une expérience de ChIP-on-chip et ChIP-seq . . . . .	33
1.12 Résolution des expériences ChIP-on-chip et ChIP-seq . . . . .	35
1.13 Expérience d'empreinte à la DNase I chez la levure : vers une résolution au nucléotide près . . . . .	36
1.14 Les différents types de CRMs et leurs marques épigénétiques . . . . .	38
1.15 Différents <i>enhancers</i> conduisent à différents patterns d'expression . . . . .	40
1.16 Deux modèles d' <i>enhancers</i> : enhanceosome et billboard . . . . .	42
1.17 L'enhanceosome de l'interferon- $\beta$ . . . . .	43
1.18 Flexibilité du code de cis-régulation au cours de l'évolution chez les <i>Drosophiles</i> .	44
1.19 Évolution de la fixation de HNF4 $\alpha$ chez les mammifères . . . . .	46
1.20 « Shadow enhancer » du gène de segmentation <i>Hunchback</i> . . . . .	48
1.21 De l' <i>enhancer</i> au super- <i>enhancer</i> . . . . .	49
1.22 Différentes approches pour la prédiction des CRMs . . . . .	51
1.23 Méthodes de validation des CRMs par transfection et transgenèse . . . . .	56
1.24 Impact physiologique de la délétion et de la mutation d'un enhancer . . . . .	57
1.25 Évolution du coût de séquençage . . . . .	59
1.26 Visualisation de données ChIP-seq via le site UCSC . . . . .	62

Liste des figures

---

1.27 Les différentes données obtenues par le projet ENCODE . . . . .	63
<b>Modèles de fixation des Facteurs de Transcription à l'ADN. . . . .</b> 65	
2.1 Différents modèles pour décrire les corrélations entre nucléotides dans les sites de fixation de facteurs de transcription . . . . .	68
2.2 Illustration d'un système dont on veut maximiser l'entropie . . . . .	72
2.3 Chaleur spécifique pour différents TFs . . . . .	79
2.4 Histogrammes des valeurs des champs $h$ et couplages $J$ . . . . .	81
 <b>Imogene : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle . . . . .</b> 83	
3.1 Illustration de l'approche Espérance-Maximisation . . . . .	85
3.2 Conditions initiales . . . . .	100
3.3 Corrélations entre les échantillons . . . . .	102
3.4 Estimation de la convergence . . . . .	104
3.5 Comparaison des approches par MCMC et par descente de gradient . . . . .	105
 <b>Étude de la différenciation épidermale chez la drosophile . . . . .</b> 107	
4.1 Frontière d'efficacité de Pareto . . . . .	110
 <b>Étude de la différenciation musculaire chez la souris . . . . .</b> 113	
5.1 Courbe ROC de MEF3 . . . . .	117
 <b>Statistiques génomiques . . . . .</b> 121	
A.1 Distribution des tailles intergéniques et introniques chez différentes espèces . .	122

---

# Principales abréviations utilisées

---

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



---

# Avant-propos

---

Cette thèse se présente sous la forme suivante...

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

**these:version du mercredi 17 juillet 2013 à 10 h 31**

*Avant-propos*

---

---

# Chapitre 1

## Introduction générale.

---

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

*Chapitre 1. Introduction générale.*

---

1.6.1	Méthodes utilisant la concentration en sites de fixation . . . . .	50
1.6.2	Méthodes utilisant la phylogénie . . . . .	52
1.6.3	Méthodes utilisant les marques épigénétiques et de ChIP-seq pour des TFs . . . . .	54
1.6.4	Validation expérimentale . . . . .	55
1.6.5	Implication des CRMs dans les maladies humaines . . . . .	57
<b>1.7</b>	<b>Bases de données . . . . .</b>	<b>59</b>
1.7.1	Obtention de données génomiques . . . . .	59
1.7.2	Obtention de données sur les TFs . . . . .	61
1.7.3	Outils de visualisation . . . . .	61
1.7.4	Le projet ENCODE . . . . .	63

## 1.1 Le phénotype cellulaire

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

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

Bien que possédant toutes le même matériel génétique (à quelques variations près), les cellules d'un organisme apparaissent d'emblée comme hétérogènes, que ce soit dans leur forme ou dans leurs constituants. Par exemple, chez l'homme, les neurones sont composés d'un corps d'une dizaine de microns de diamètre contenant le noyau, de milliers de prolongements dont certains peuvent atteindre une taille de plus de 1.5 mètre, et sont riches en ions sodium et potassium, tandis que les fibres musculaires squelettiques sont de forme longue et tubulaire, possèdent plusieurs dizaines de noyaux et expriment actine et myosine.

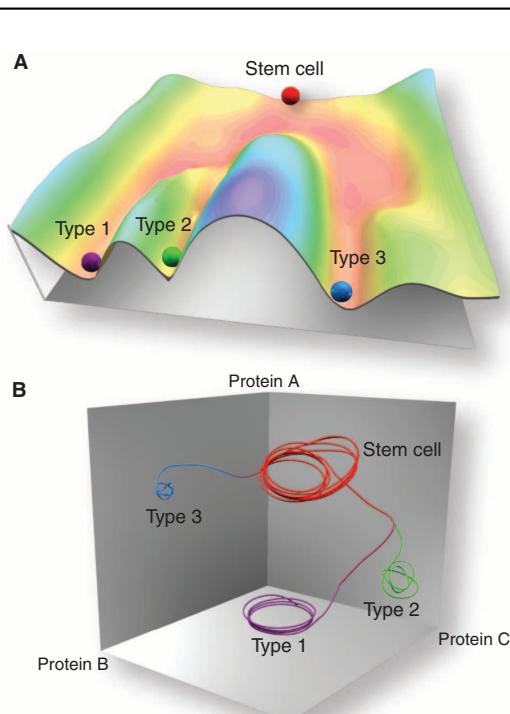
Cette diversité semble néanmoins limitée. Aussi, parmi les  $\sim 6 \cdot 10^{13}$  cellules du corps humain, on peut distinguer  $\sim 320$  différents types cellulaires ([Brazma et al., 2001](#)). Bien entendu, ce nombre dépend du seuil de similarité choisi : deux cellules d'un même type n'expriment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

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

---

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

### 1.1.2 La différenciation cellulaire



**FIGURE 1.1 – Le paysage de la différenciation cellulaire.**

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

L'acquisition d'un phénotype cellulaire particulier au sein d'un organisme est le sujet de la biologie du développement. Cette acquisition passe par différentes étapes de différenciation cellulaire. Schématiquement, au cours du développement d'un organisme, des cellules non différenciées ou souches empruntent un chemin unidirectionnel de différenciation qui restreint peu à peu le nombre de types cellulaires qu'elles peuvent potentiellement devenir, passant de l'état souche totipotent à des états pluripotents successifs avant la différenciation

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

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

La notion de trajectoire de différenciation peut être rendue plus parlante en adoptant une représentation de système dynamique. Comme nous l'avons vu en 1.1.1, la cellule contient de nombreux composants : gènes, protéines ou autres métabolites, qui pris dans leur ensemble déterminent à un instant donné l'état cellulaire. Il est ainsi possible de représenter l'état cellulaire à un temps donné comme un point dans un espace de grande dimension dans lequel chaque axe représente l'abondance d'un certain composant (fig 1.1B). De par leur rôle primordial dans la définition de l'état cellulaire, l'expression des protéines (et donc des gènes qui les produisent) domine généralement ces composants, et on parle de « niveau d'expression génétique » pour décrire leur abondance. Les changements d'expression génétique, au cours desquels certains gènes vont être activés et d'autres réprimés, induisent un changement de l'état cellulaire, ce qui se traduit par une trajectoire dans l'espace des états. Ces changements d'expression restreignent finalement l'état cellulaire à une certaine région, définie comme un « attracteur » de la dynamique. Une fois au sein d'un attracteur, l'état cellulaire est robuste aux perturbations du niveau d'expression génétique des différentes composantes. Les attracteurs peuvent alors être vu comme des types cellulaires distincts correspondant aux différentes vallées de la représentation de Waddington ([Kaufmann, 1993](#)).

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



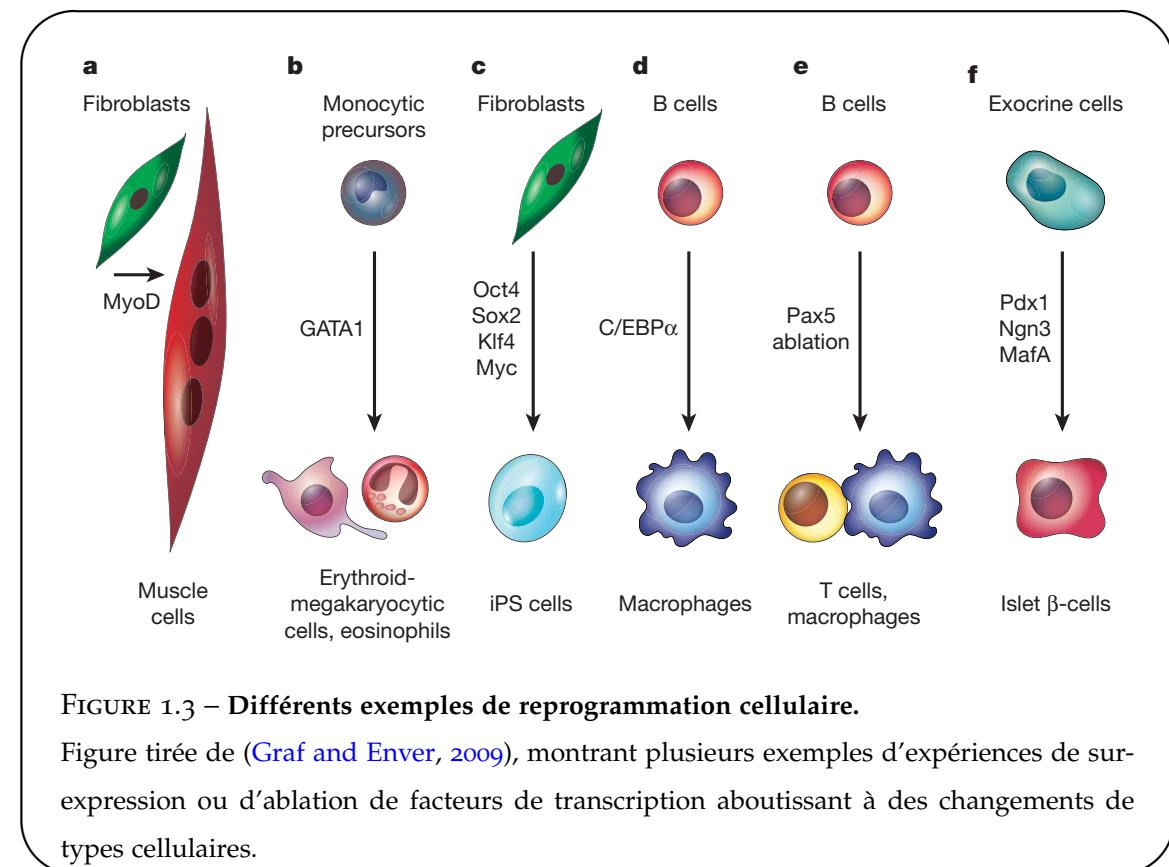
FIGURE 1.2 – Spécification spatio-temporelle du type cellulaire.

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

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

### 1.1.4 La reprogrammation cellulaire

Depuis plusieurs décennies, différentes expériences ont exhibé la plasticité des états différenciés, élargissant ainsi considérablement la vision classique selon laquelle des cellules souches totipotentes se différencient de manière irréversible en des cellules de moins en moins plastiques, jusqu'à atteindre un état différencié stable. Par exemple, (Blau et al., 1985) ont mon-



tré en 1985 que des programmes d’expression génétique dormants peuvent être exprimés de manière dominante dans des cellules différenciées par la fusion de différents types cellulaires : ainsi, la fusion de cellules musculaires avec des cellules non musculaires permettait l’activation des gènes de type musculaire dans le type cellulaire non musculaire. Puis différents travaux ont montré qu’il était possible de convertir des lignées de cellules différenciées en un autre type cellulaire en introduisant certaines protéines régulatrices de la transcription, ou Facteurs de Transcription (TFs) ([Davis et al., 1987](#); [Kulessa et al., 1995](#)) : on parle alors de trans-différenciation, dont l’un des exemples canoniques est la différenciation de cellules de la peau ou fibroblastes en cellules musculaires par l’introduction du facteur de différenciation myogénique MyoD (voir fig 1.3). Parallèlement, des expériences réalisées chez plusieurs espèces de mammifères ont montré que le transfert de noyaux de cellules différenciées embryonnaires ou adultes dans un oeuf énucléé peut mener à la formation d’un organisme complet, montrant de manière univoque que l’identité des cellules différenciées 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

## Chapitre 1. Introduction générale.

---

cellules souches puripotentes par simple introduction d'un « cocktail » de 4 facteurs de transcription : Oct4, Sox2, c-Myc et Klf4 ([Takahashi and Yamanaka, 2006](#)) (fig 1.3C).

## 1.2 Les réseaux de régulation génétique

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

### 1.2.1 Vision cybernétique de la cellule

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

### 1.2.2 Divers modes de régulation

Les modes de régulation qui permettent à la cellule d'interpréter des signaux afin de changer d'état sont nombreux. Nous allons nous concentrer ici sur ceux affectant la production d'ARNs ou de protéines (fig. 1.5).

- **Régulation génétique**

Tout d'abord, un réseau d'expression génétique est caractérisé par un jeu d'interactions entre différents gènes. Ici, nous centrons notre attention sur les gènes codant pour des pro-

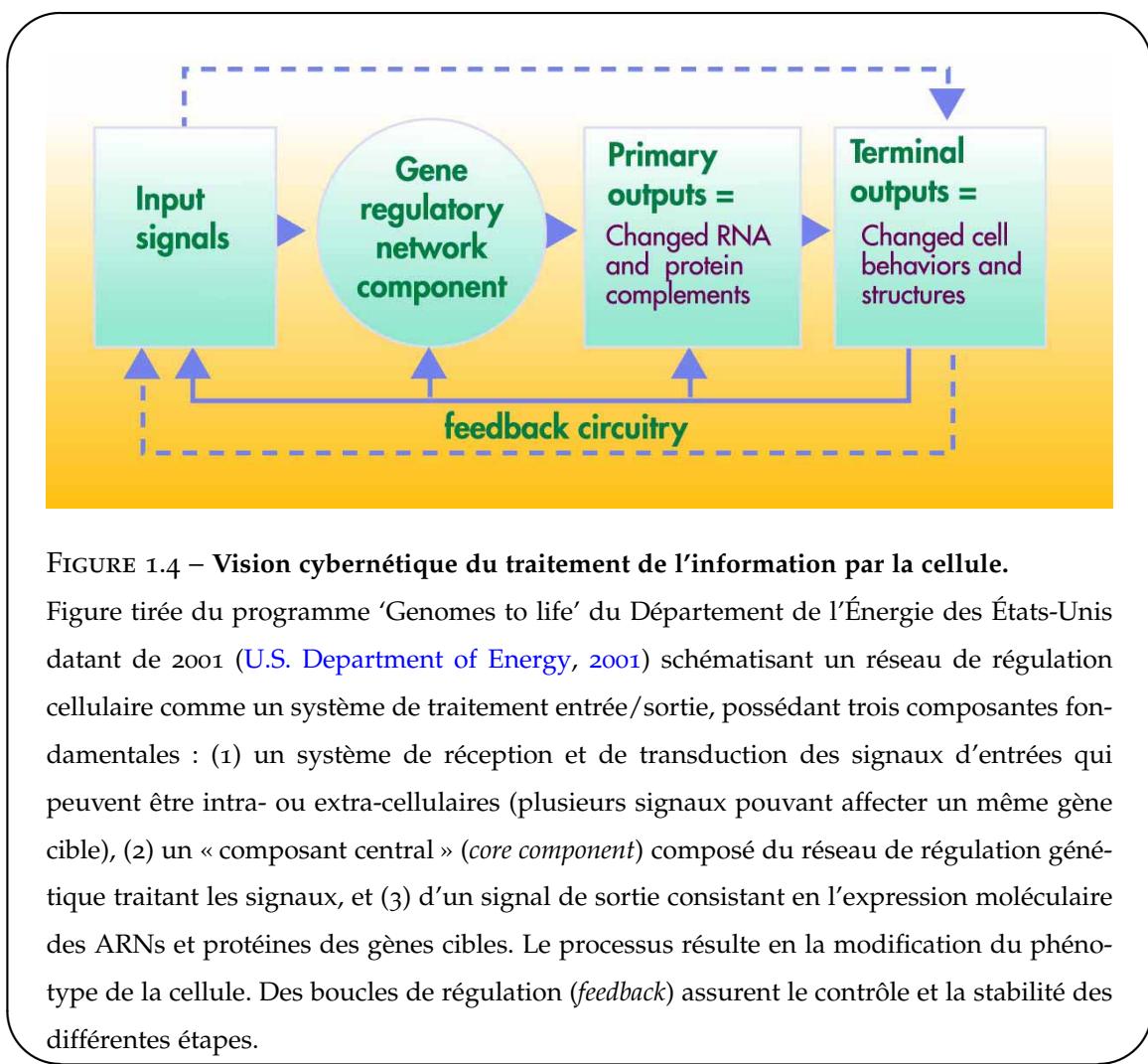
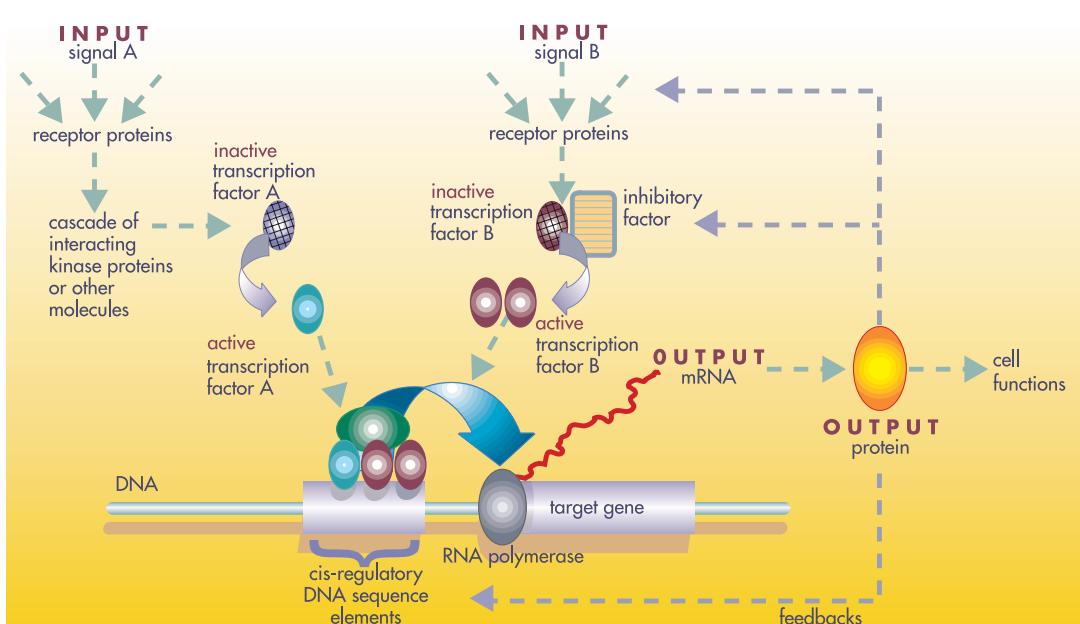


FIGURE 1.4 – Vision cybernétique du traitement de l’information par la cellule.

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



**FIGURE 1.5 – Un réseau de régulation génétique type.**

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

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

### • Régulation épigénétique

Outre la régulation génétique, due à l'action de protéines issues de séquences codantes et se fixant sur des séquences d'ADN – régulation qui est donc entièrement encodée dans le génome et transmise à la descendance –, il existe un autre mode de régulation de la transcription des gènes qui permet notamment d'acquérir une modification d'expression génétique transmise à la descendance sans qu'il y ait modification du code génétique : c'est la régulation épigénétique. Cette régulation passe notamment par la modification des propriétés chimiques de l'ADN et des histones sur lequel il s'enroule pour former la chromatine (fig. 1.6)<sup>2</sup>. Ainsi, la méthylation des dimères CpG de l'ADN<sup>3</sup> au niveau des régions riches en CG, ou îlots CpG, situées près de nombreux promoteurs et habituellement dépourvues de ces marques conduit à une inactivation du gène cible (Bird, 2002). Par ailleurs, la méthylation des histones au niveau des résidus lysines entraîne la fermeture de la chromatine, empêchant l'expression du ou des

2. Il est à noter que certains emploient le terme épigénétique pour qualifier la fixation des TFs sur l'ADN. Ici, le terme épigénétique réfère seulement aux modifications chimiques affectant les histones et l'ADN, et donc l'accessibilité du génome

3. 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.

---

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 détail en section [1.5.1](#).

### • Régulation post-transcriptionnelle

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

### • Régulation post-traductionnelle

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

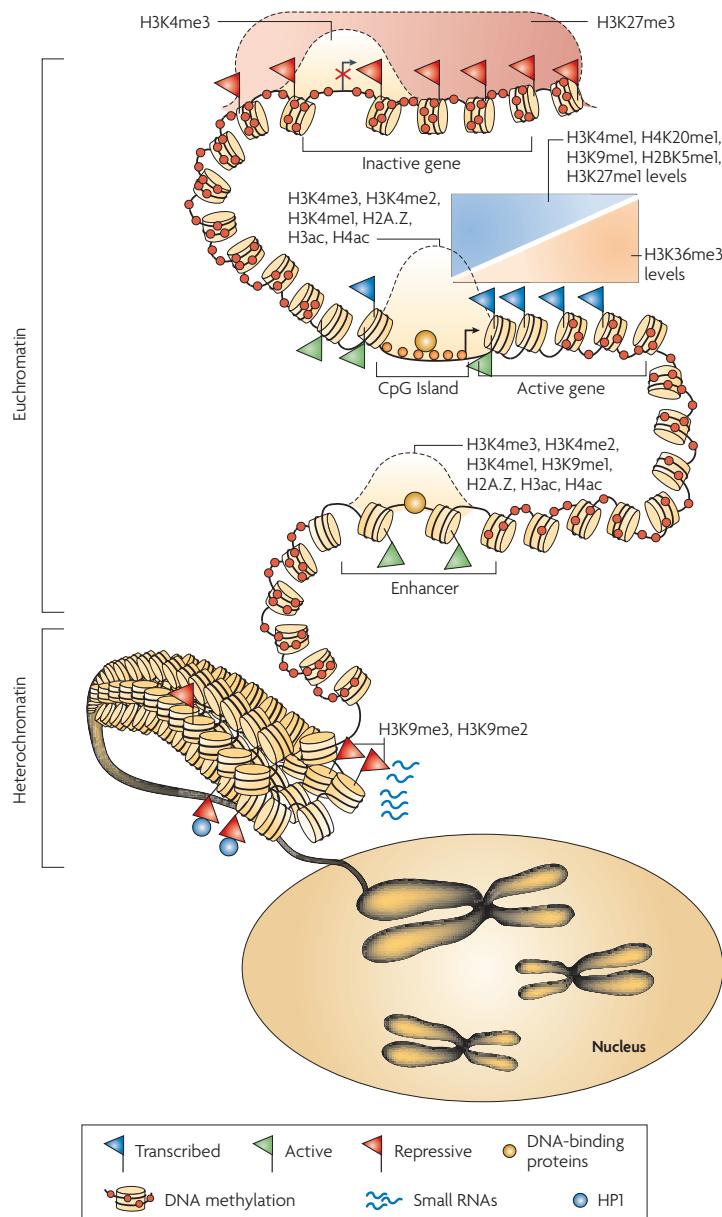


FIGURE 1.6 – Caractéristiques de l'épigénome.

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

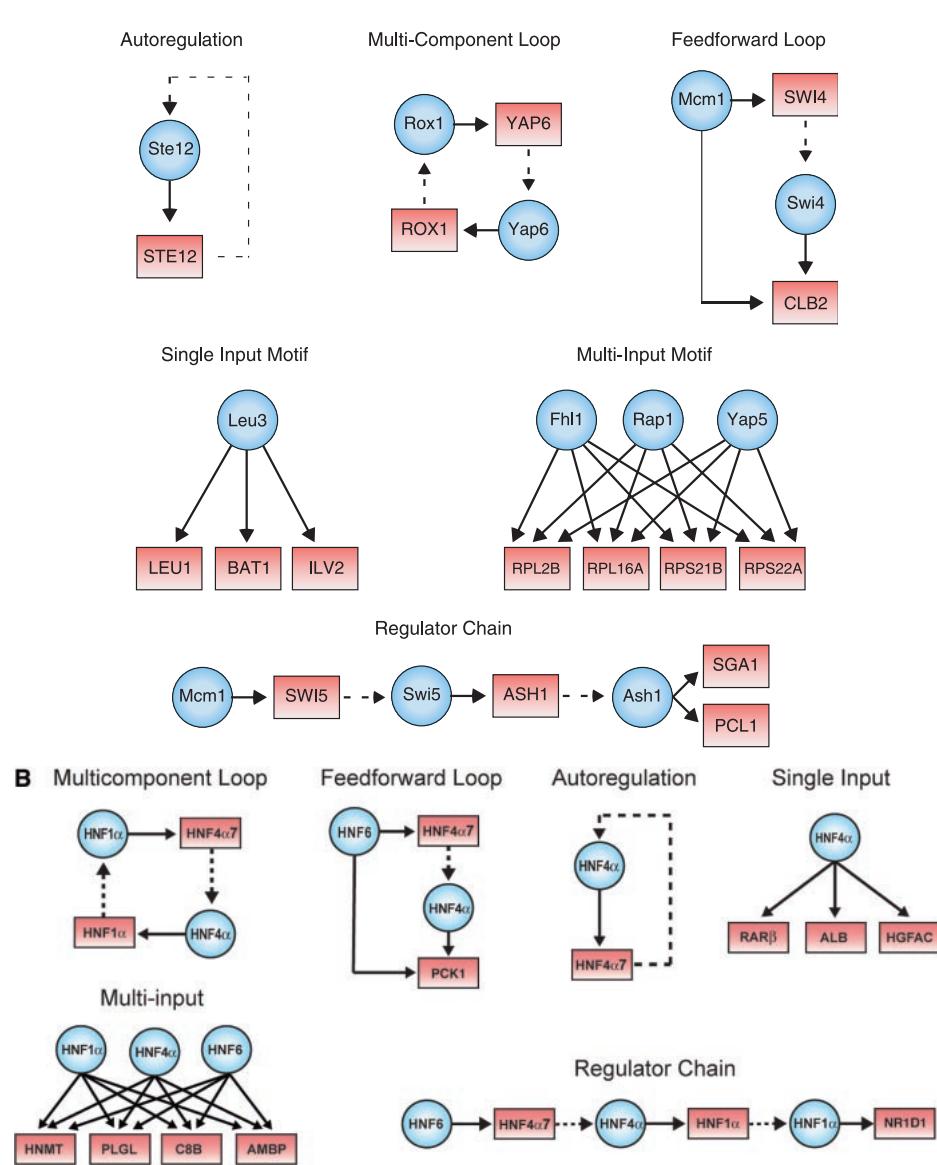


FIGURE 1.7 – Exemples de motifs dans les réseaux de régulation génétique.

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

### 1.2.3 Câblage du réseau et fonction

Maintenant que nous avons vu la nature des interactions au sein des réseaux génétiques, nous pouvons nous pencher sur leur structure. Notamment, plusieurs études réalisées chez divers organismes de la bactérie à l'homme ont révélé que les réseaux de transcription contiennent un petit ensemble de motifs de régulation récurrents, appelés motifs de réseaux (Alon, 2007a; Shen-Orr et al., 2002; Milo et al., 2002) (fig. 1.7). 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). Une explication possible de la récurrence de ces motifs est liée aux fonctions qu'ils remplissent. Par exemple, la boucle d'autorégulation négative, qui est trouvée chez la moitié des répresseurs d'*Escherichia coli*, possède deux fonctions : l'une est de parvenir rapidement à un état d'équilibre en utilisant un promoteur fort, l'autre est de servir de tampon au bruit d'expression (Alon, 2007b). Un autre motif récurrent est la boucle feedforward. Celle-ci consiste en 3 gènes : un régulateur X, qui régule Y, tous deux régulant Z. Dans le cas où des interactions sont des activations et que X et Y sont requis pour activer Z, cette boucle peut servir de tampon au bruit d'expression de X, évitant que des fluctuations de son niveau d'expression n'entraîne par erreur l'activation de Z.

### 1.2.4 Évolution des réseaux génétiques

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

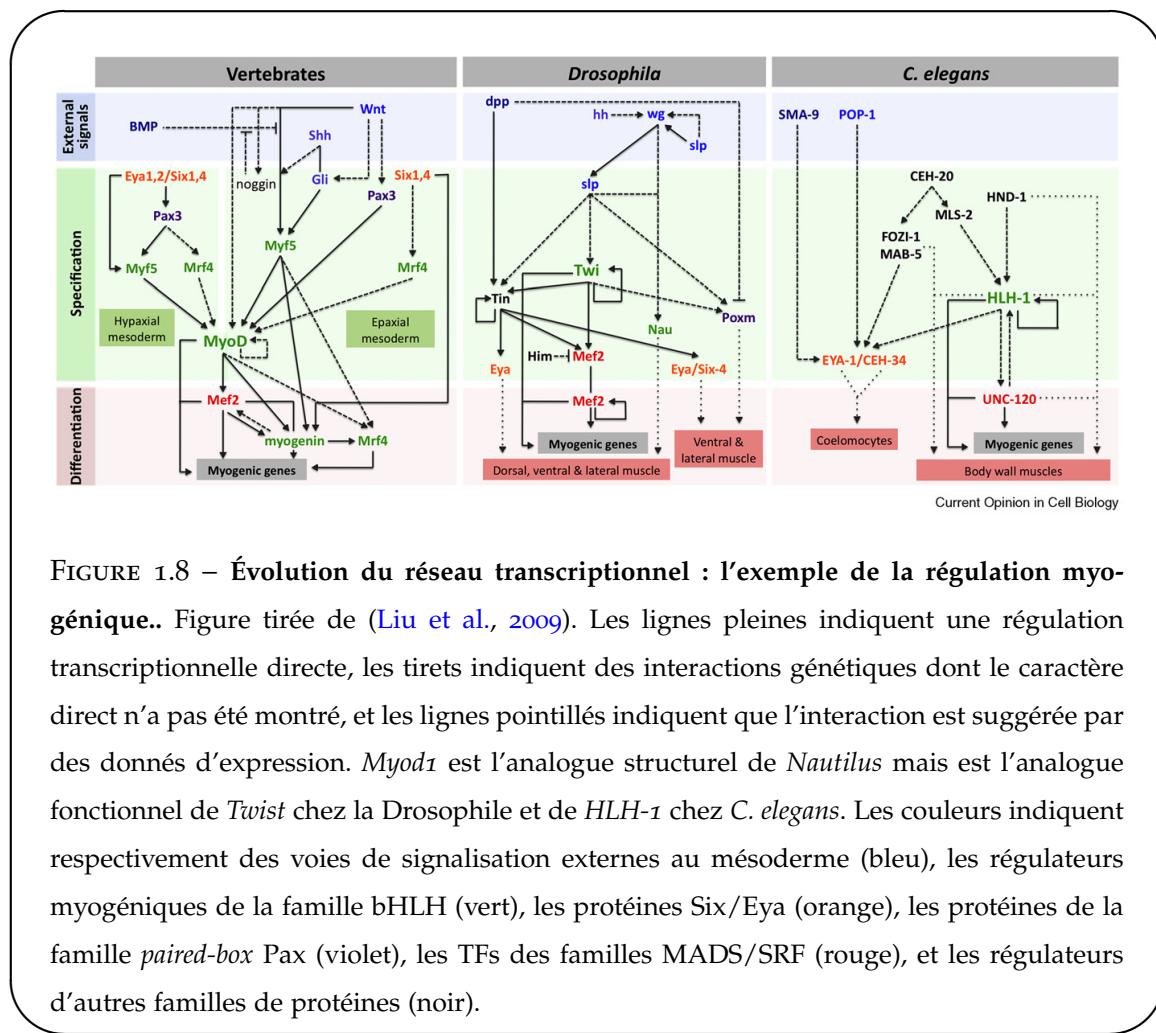


FIGURE 1.8 – Évolution du réseau transcriptionnel : l'exemple de la régulation myogénique.. Figure tirée de (Liu et al., 2009). Les lignes pleines indiquent une régulation transcriptionnelle directe, les tirets indiquent des interactions génétiques dont le caractère direct n'a pas été montré, et les lignes pointillés indiquent que l'interaction est suggérée par des données d'expression. *Myod1* est l'analogie structurelle de *Nautilus* mais est l'analogie fonctionnelle 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).

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

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

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

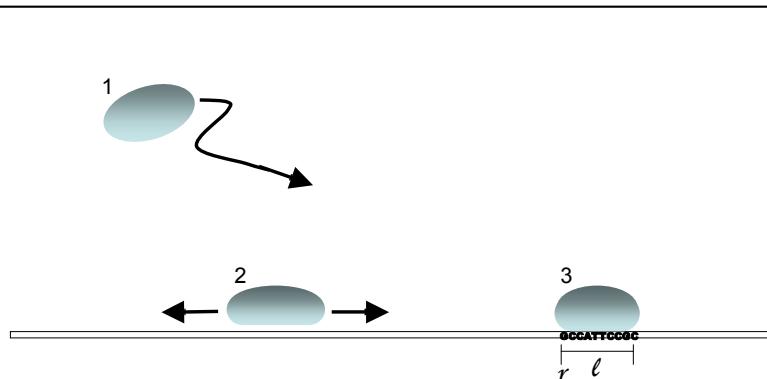
Nous voyons donc que l'évolution d'un réseau génétique possède de multiples facettes : conservation de motifs de réseau fonctionnellement importants (dans notre exemple, la boucle feedforward au cœur du réseau régissant l'aspect temporel de l'expression des cibles), recâblage des interactions pour traiter différents signaux d'entrée... Par ailleurs, il apparaît que

## Chapitre 1. Introduction générale.

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

### 1.3 Les interactions protéine-ADN : modèles mathématiques

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



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

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

### 1.3.1 Modes de recherche du site de fixation par le TF

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

### 1.3.2 Modèle PWM

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

## Chapitre 1. Introduction générale.

---

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

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

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

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

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

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

où  $L$  est la taille du site de fixation et  $\pi_b$  correspond à la probabilité *a priori* d'observer la base  $b$  dans le génome. Il est usuel de définir l'énergie relativement au fond génomique :

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

L'énergie totale d'un site  $S_i$  est alors

## 1.3. Les interactions protéine-ADN : modèles mathématiques

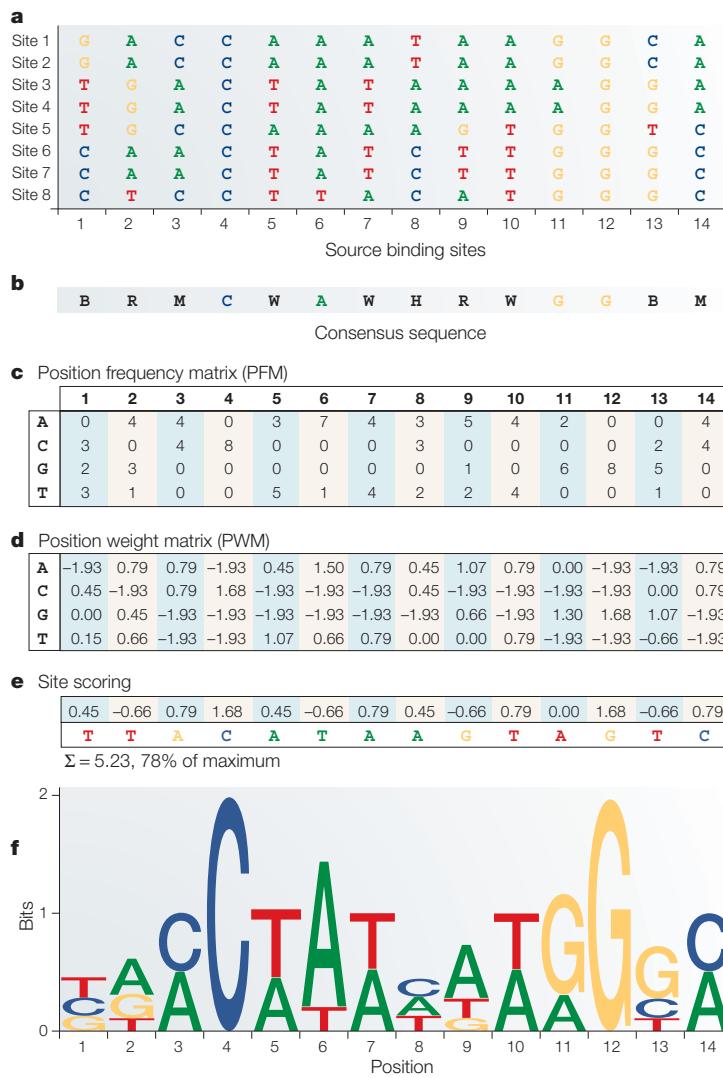


FIGURE 1.10 – Construction et utilisation du modèle PWM. Figure tirée de (Wasserman and Sandelin, 2004). (a) Supposons connus un certain nombre de sites de fixation d'un facteur de transcription (dans ce cas MEF2). (b) Séquence consensus correspondante utilisant les symboles IUPAC. (c) Une matrice de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

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

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

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

### 1.3.3 Modèle biophysique

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

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



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

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

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

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

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

### 1.3.4 Modèle thermodynamique

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

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

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

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

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

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

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

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

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

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

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

---

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


---

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

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

En utilisant à nouveau la définition de  $E_i$  en éq.[1.10](#), la probabilité de fixation d'un site à la position  $i$  s'écrit

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

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

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

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

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

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

## 1.4 Les interactions protéine-ADN : mesures expérimentales

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

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

- **Approche microfluidique : MITOMI**

En 2007, 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 finement contrôlé. Le fluide contient des gènes synthétiques codant pour le TF ainsi que du matériel permettant la synthèse de la protéine directement au sein de la chambre, ce qui évite de purifier préalablement le TF. Chaque chambre du système contient des anticorps fixés à la surface permettant de capturer le TF et une certaine concentration d'une séquence d'ADN spécifique contenant une marque fluorescente. Le système contient ainsi des centaines de séquences d'ADN différentes, chacune étant présente à différentes concentrations. Lorsque le TF est fixé par les anticorps, il recrute des séquences d'ADN selon leur affinité. Celles qui ne se fixent pas sont lavées. Au final, les séquences fixées produisent un signal de fluorescence. La comparaison des signaux pour différentes concentrations d'ADN donne accès au rapport des constantes d'équilibre  $K_{eq}$  (eq. 1.6). La comparaison avec une séquence référence dont la constante  $K_{eq}$  est connue permet alors de déterminer le  $K_{eq}$  absolu pour chaque séquence de fixation.

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

---

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


---

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

- **Approche physique : la microscopie SPR**

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

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

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

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

## Chapitre 1. Introduction générale.

---

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

### • Approche par purification des séquences fixées : SELEX et HT-SELEX

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

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

---

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


---

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

#### 1.4.2 Approche clonale : la technique de simple hybride

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

## Chapitre 1. Introduction générale.

---

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

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

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

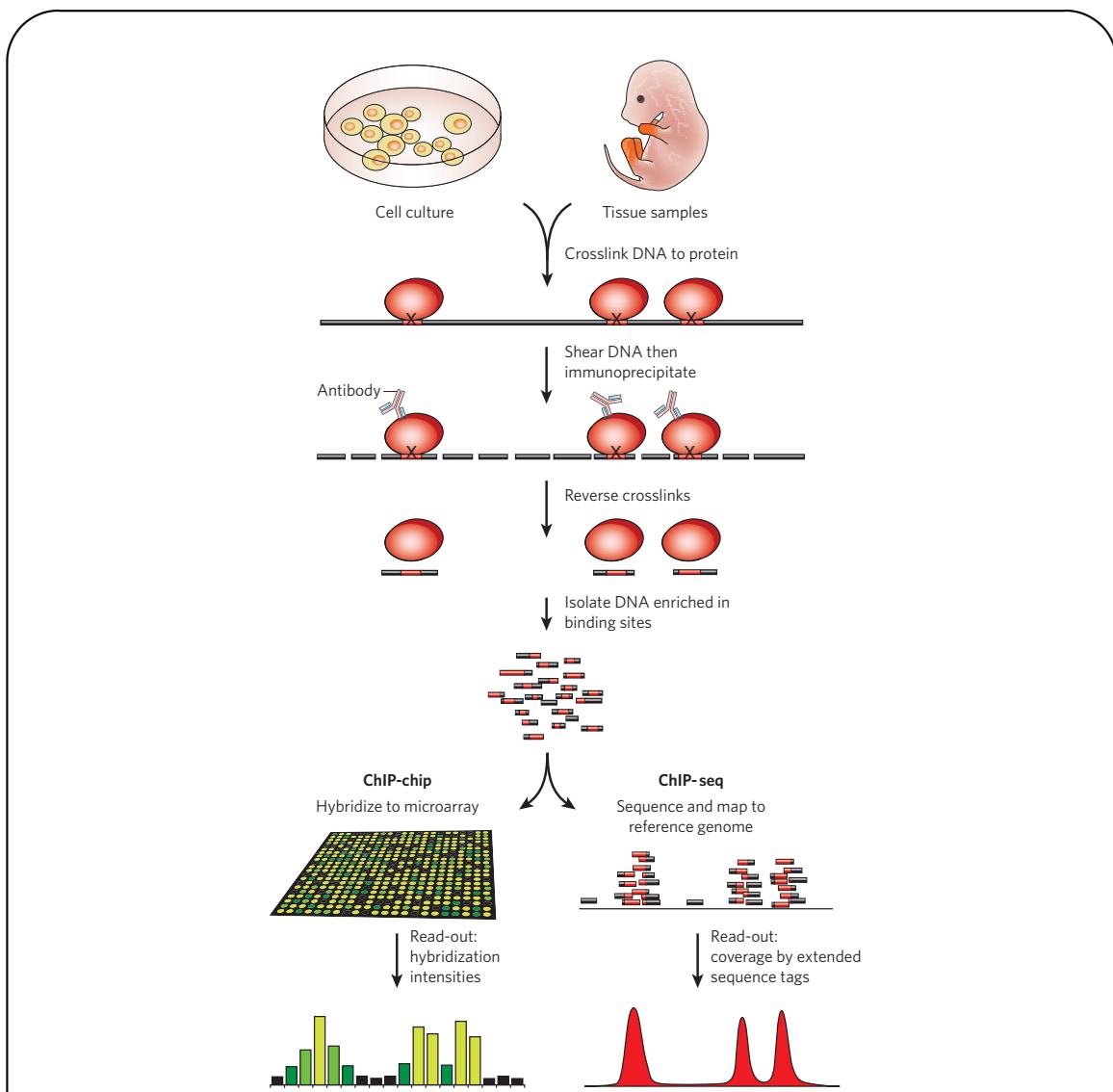
### 1.4.3 Approches *in vivo* : ChIP-on-chip, ChIP-seq, DNase I

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

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

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

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



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

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

## Chapitre 1. Introduction générale.

---

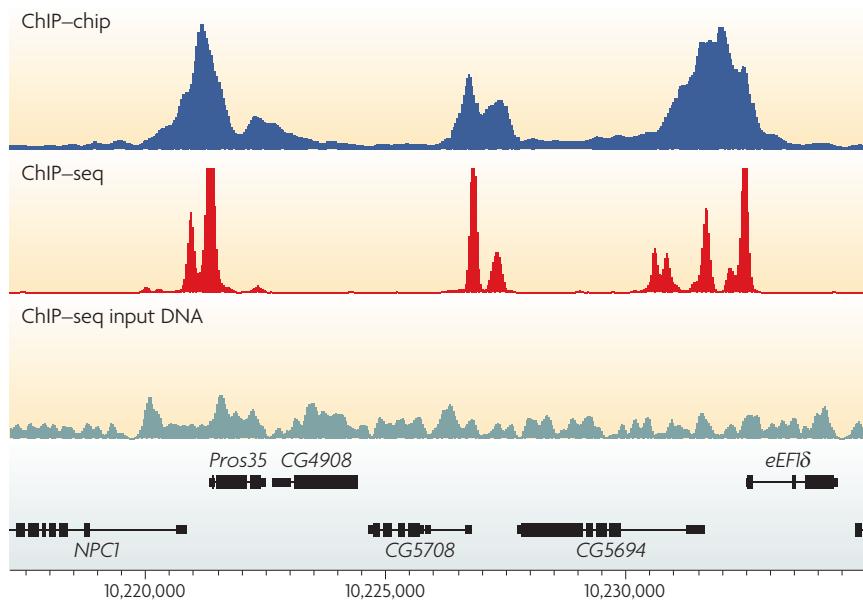
protéine-ADN en liaisons covalentes. Une fois les protéines fixées, la chromatine est découpée par digestion enzymatique ou en la soumettant à des ultrasons (c'est la sonication), résultant en des fragments de taille variant entre 200 et 600bp. Ces fragments sont ensuite immunoprécipités en présence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

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

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

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

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



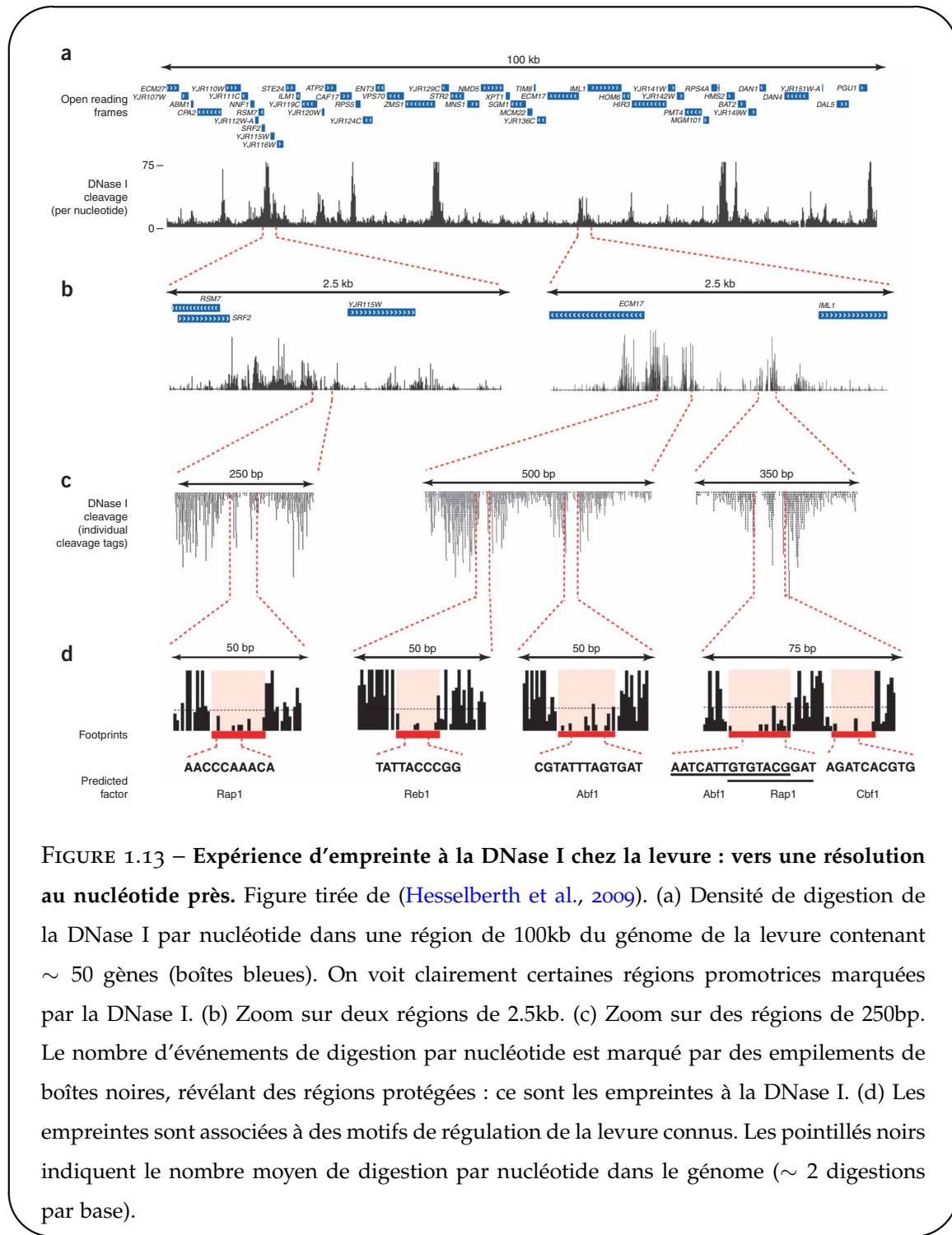
**FIGURE 1.12 – Résolution des expériences ChIP-on-chip et ChIP-seq.** Figure tirée de (Park, 2009), montrant les profils de fixation de la protéine Chromator générés à partir d'expériences de ChIP-on-chip (intensité relative par rapport au contrôle, bleu) et de ChIP-seq (densité de séquences, rouge)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

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

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

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

## Chapitre 1. Introduction générale.



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

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

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

## 1.5 Les modules de cis-régulation (CRMs)

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

### 1.5.1 Les différents types de CRMs

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

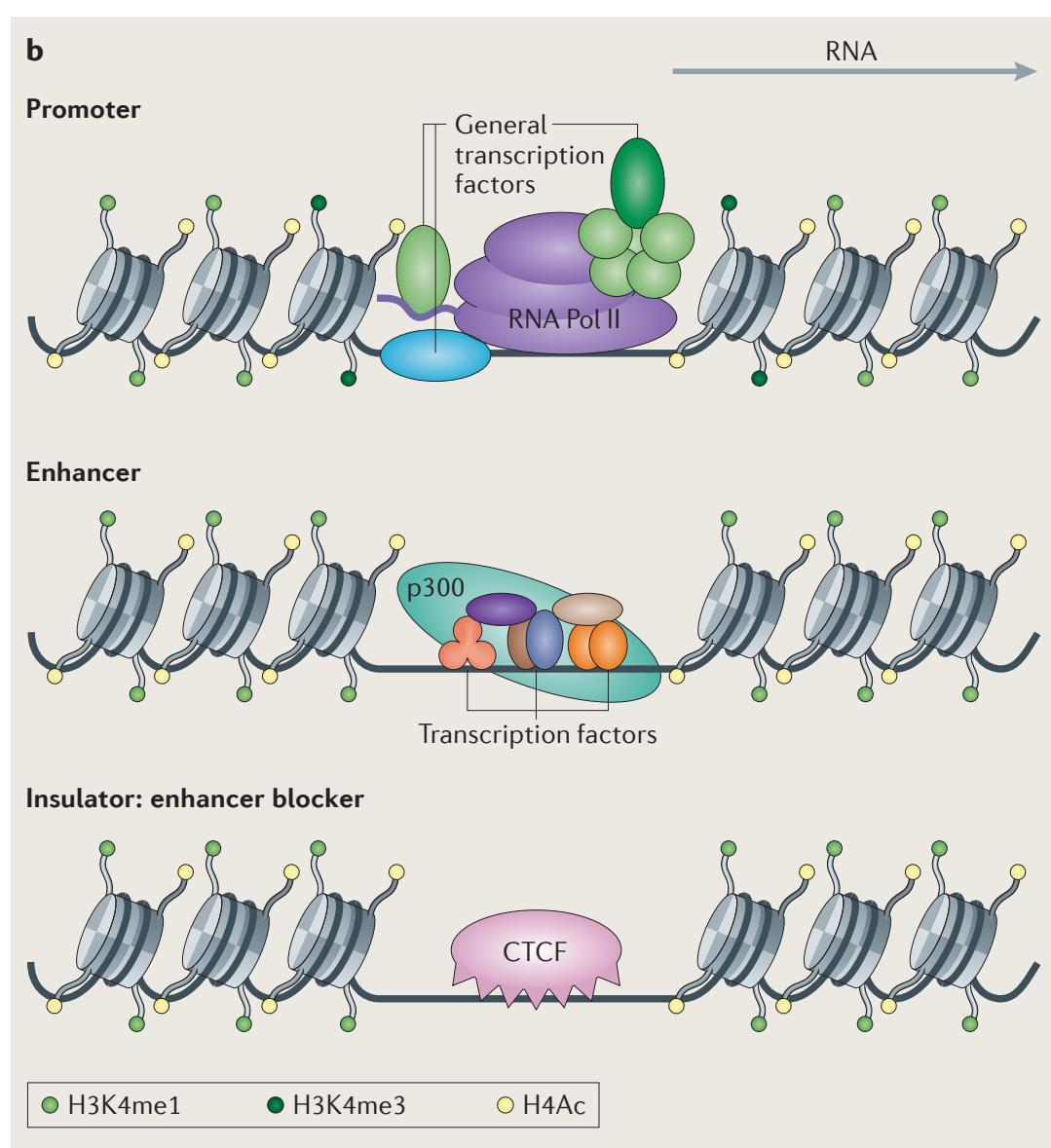


FIGURE 1.14 – Les différents types de CRMs et leurs marques épigénétiques.

Figure tirée de ([Hardison and Taylor, 2012](#)). La notion de CRM renvoie à un regroupement de sites de liaison pour un ou plusieurs facteurs de transcription. Les CRMs peuvent être regroupés en plusieurs classes : les promoteurs, les *enhancers/silencers*, et les insulateurs. Les CRMs des différentes classes partagent les marques d'acétylation H<sub>3</sub>Ac et H<sub>4</sub>Ac, les promoteurs actifs sont spécifiquement marqués par H<sub>3</sub>K4me3, et les enhancers et insulateurs par H<sub>3</sub>K4me1. Les enhancers sont par ailleurs souvent fixés par le co-activateur p300. Enfin, chez les mammifères les insulateurs recrutent CTCF pour bloquer l'activation par les enhancers.

- **Promoteurs**

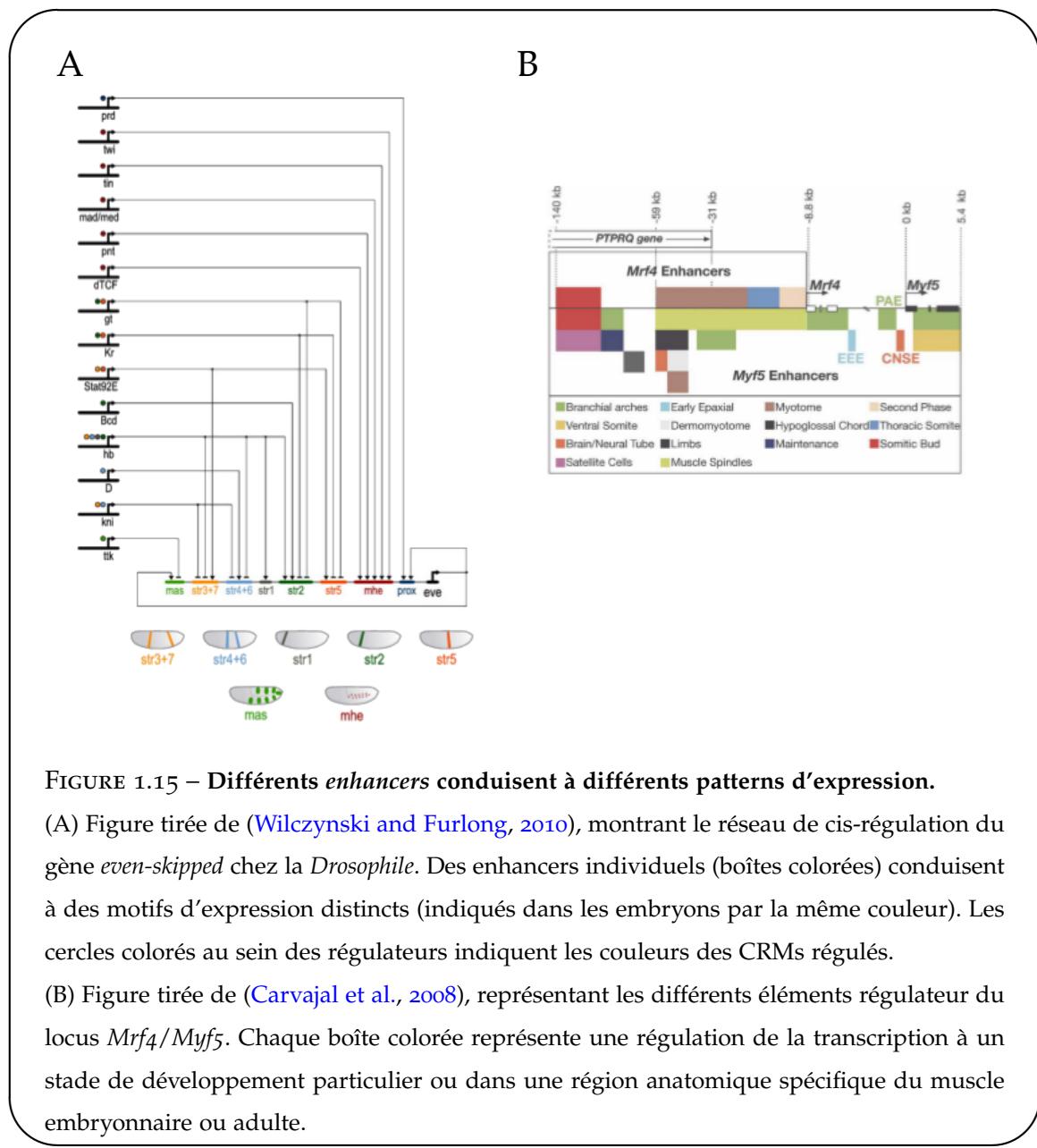
Les promoteurs permettent la fixation de l'ARN polymérase pour débuter la formation d'un transcript ARN au site d'initiation de transcription (*Transcription Start Site* ou TSS). Dans les promoteurs fixant l'ARN polymérase II (la majorité des promoteurs eucaryotes), des facteurs de transcription généraux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

- ***Enhancers et silencers***

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

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

## Chapitre 1. Introduction générale.



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

Enfin, comme décrit en fig. 1.14, les enhancers sont associés à de hauts niveaux de marque épigénétique H<sub>3</sub>K4me1 (Heintzman et al., 2009) et sont souvent fixés par le co-activateur p300 (Wang et al., 2005; Heintzman et al., 2009).

- **Insulateurs**

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

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

### 1.5.2 Grammaire des enhancers : enhanceosome vs billboard

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

- **Le modèle “enhanceosome”**

Dans ce modèle, l'architecture des sites de liaison est de première importance. Le paradigme en est l'enhanсer du gène humain interferon- $\beta$ , sur lequel 8 TFs se lient pour former

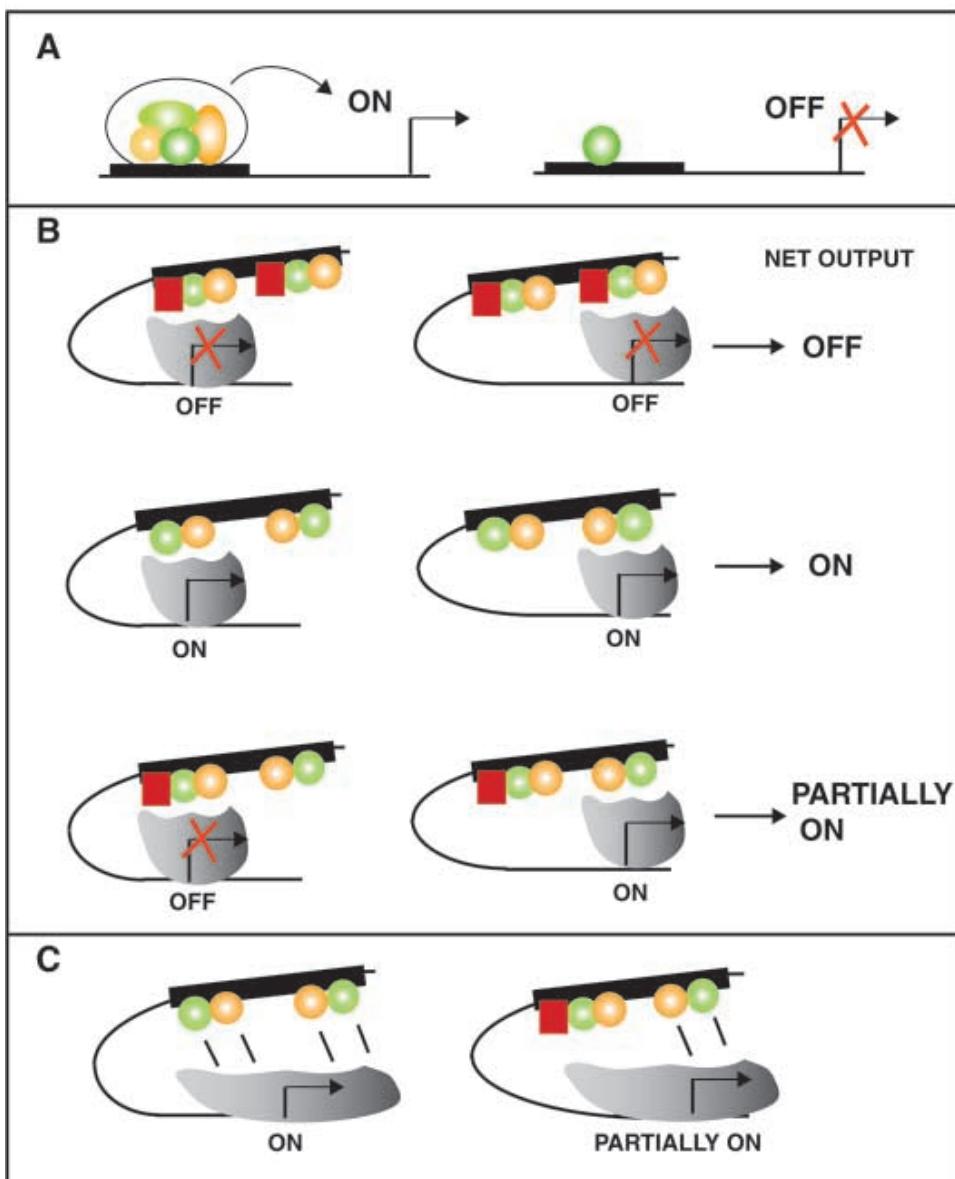
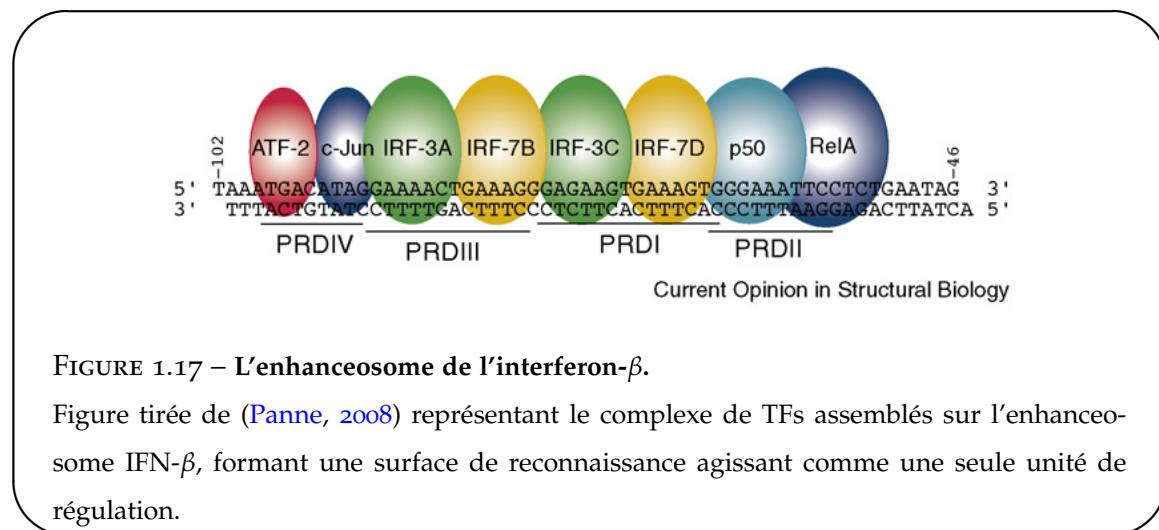


FIGURE 1.16 – Deux modèles d’enhancers : enhanceosome et billboard.

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



**FIGURE 1.17 – L’enhanceosome de l’interferon- $\beta$ .**

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

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. [1.17](#)).

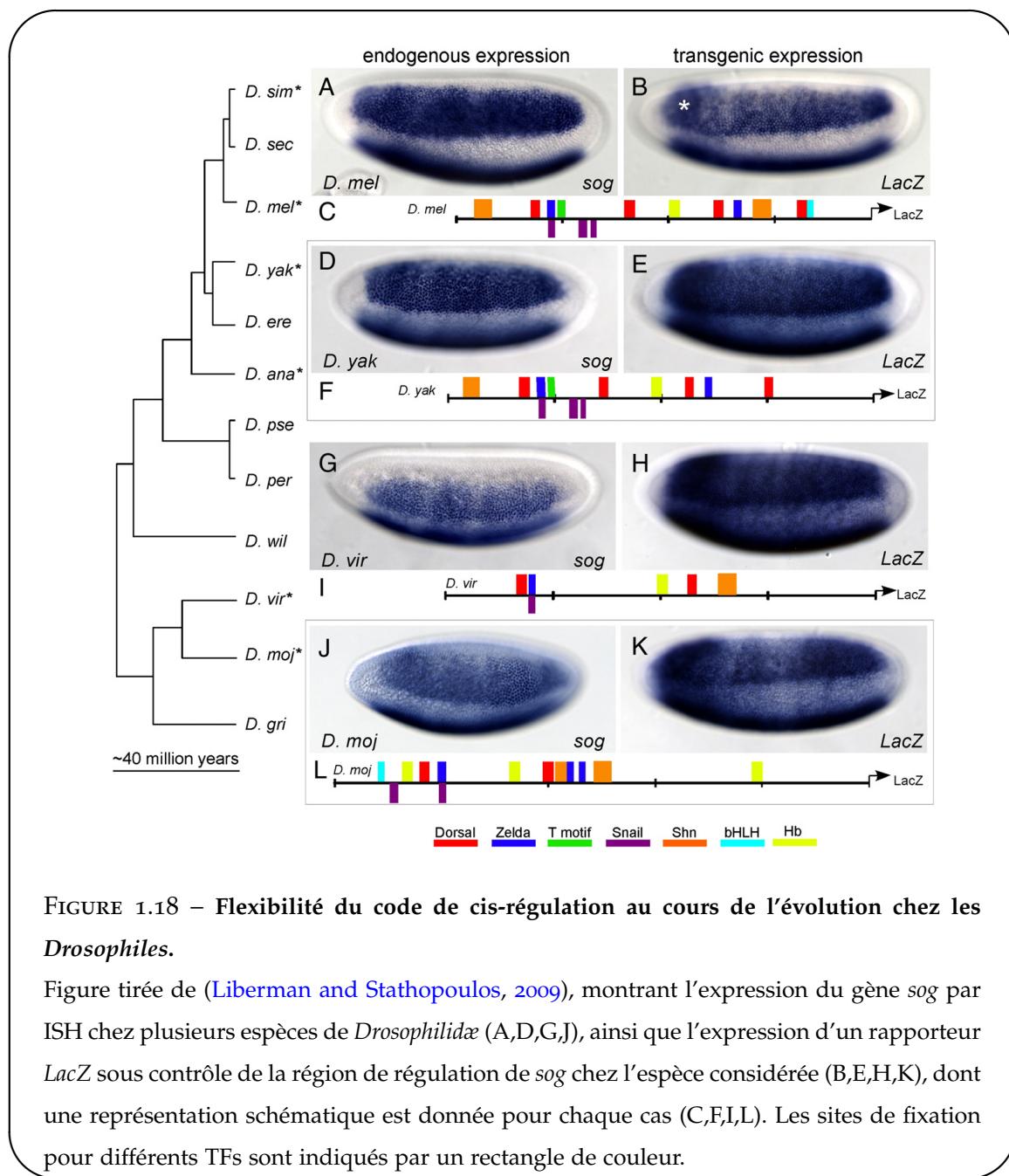
- **Le modèle “billboard”**

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

### 1.5.3 Évolution des enhancers

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

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



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

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

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

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

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

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

## Chapitre 1. Introduction générale.

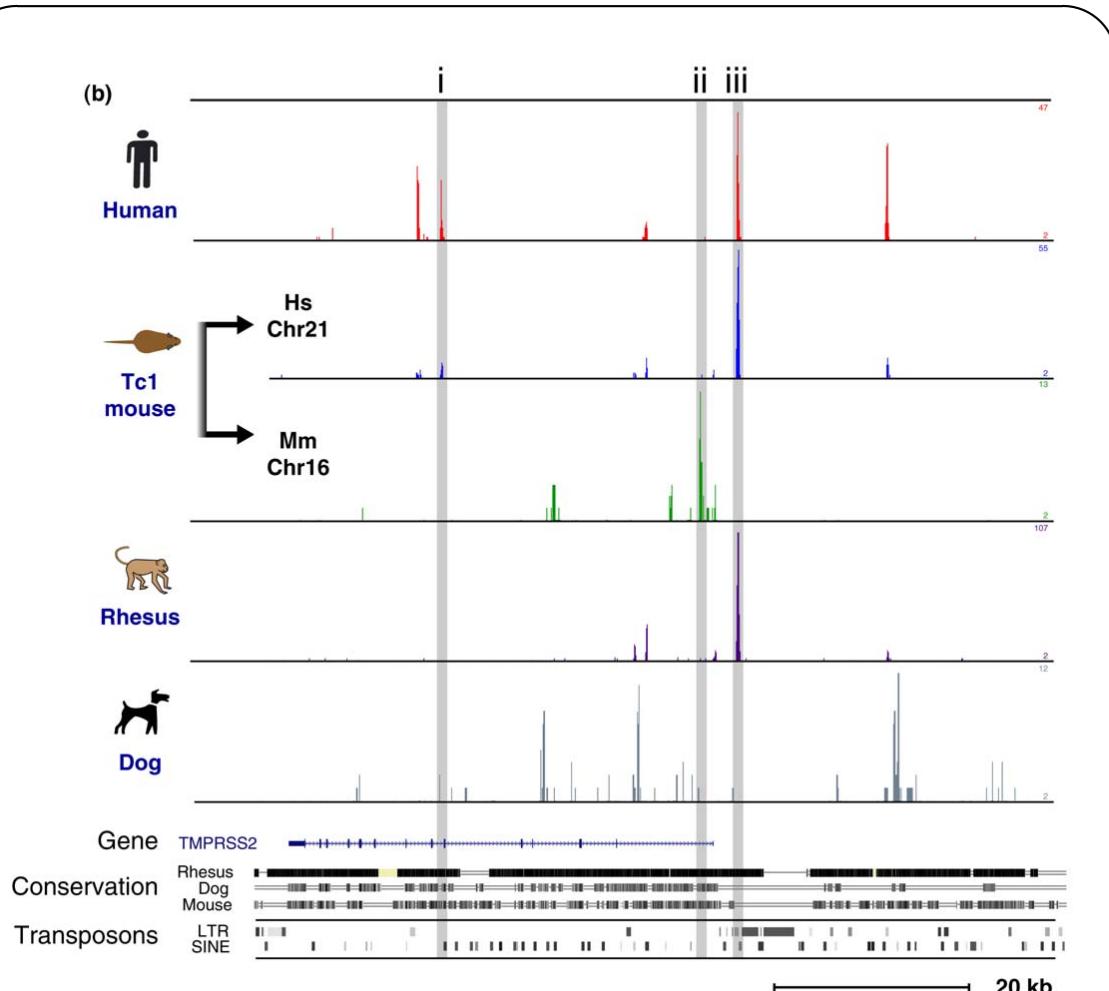


FIGURE 1.19 – Évolution de la fixation de HNF4α chez les mammifères.

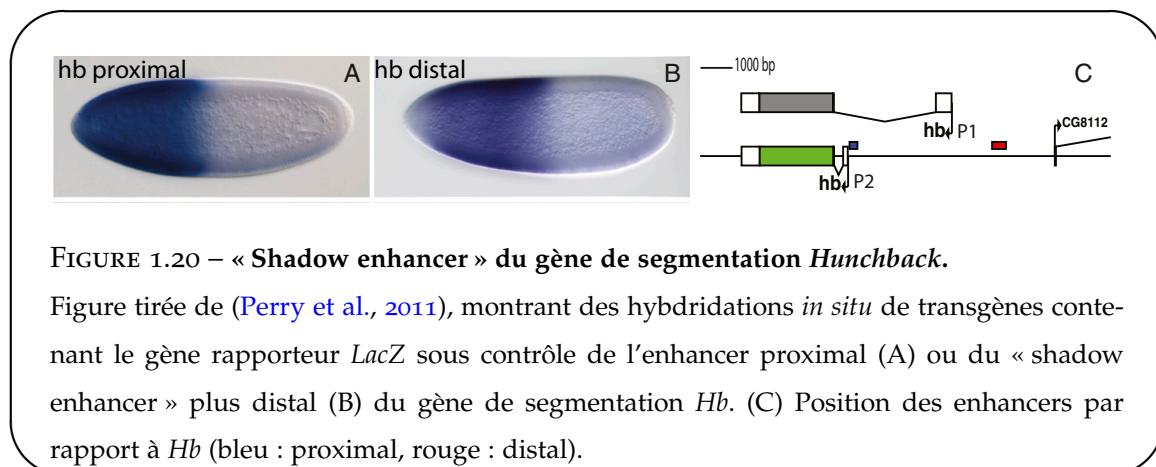
Figure tirée de (Wilson and Odom, 2009), représentant la fixation par ChIP-seq (pics de couleur) du facteur de transcription humain HNF4α (ou son homologue murin HNF4a) chez l'homme, la souris, le macaque, le chien, ainsi qu'une souris transgénique contenant le chromosome 21 humain. Les zones grisées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

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

#### 1.5.4 Les « shadow enhancers »

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

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



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

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

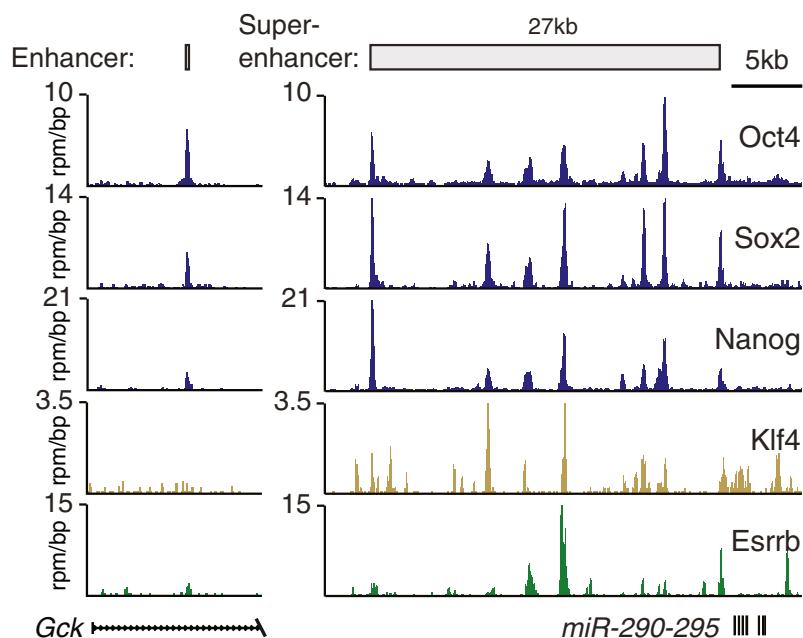
température usuelles, mais requis lorsque les embryons se développent dans des conditions de température extrêmes ([Frankel et al., 2010](#)).

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

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

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

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



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

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

## 1.6 Prédiction et validation des CRMs

### 1.6.1 Méthodes utilisant la concentration en sites de fixation

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

- **Approches utilisant des motifs connus**

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

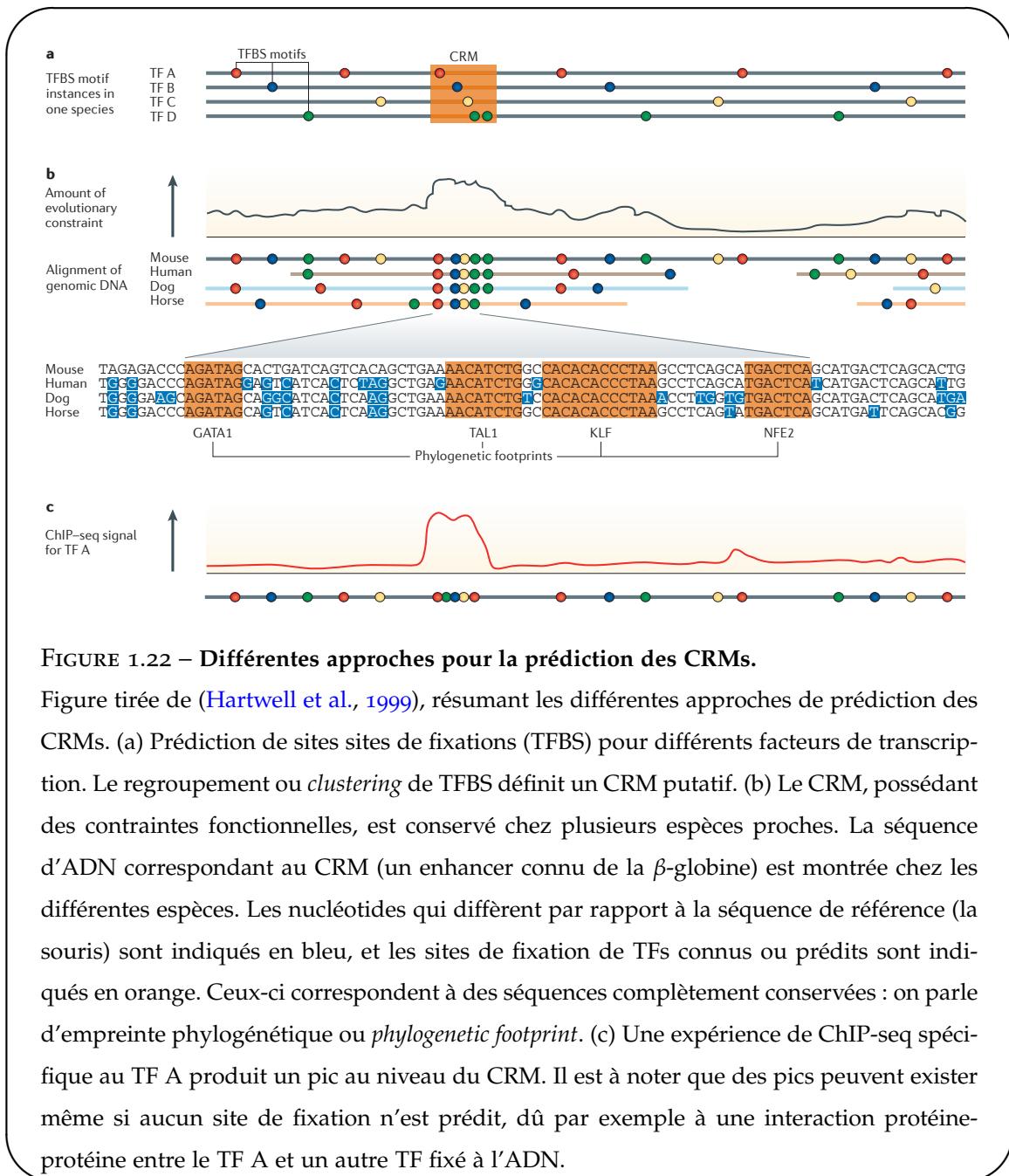


FIGURE 1.22 – Différentes approches pour la prédition des CRMs.

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

## Chapitre 1. Introduction générale.

---

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

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

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

### 1.6.2 Méthodes utilisant la phylogénie

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

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

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

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

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

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

## Chapitre 1. Introduction générale.

---

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

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

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

De même que précédemment, tous les motifs ne sont pas connus et il peut être utile d'avoir recours à de l'apprentissage direct à partir de séquences fonctionnelles connues pour aider à la prédiction. Par exemple, l'algorithme ESPER cherche des patterns (TFBS, %GC, etc) surreprésentés dans des alignements multi-espèces de CRMs connus par rapport à des alignements d'ADN *a priori* non fonctionnel ([Taylor et al., 2006](#)). Cette méthode n'est pas restreinte à l'analyse de séquences conservées puisqu'elle peut potentiellement capturer des signatures de changements systématiques. La prédiction de régions de haut potentiel de régulation recouvre presque entièrement les prédictions de PReMods, et le test par transfection de ces régions à proximité de gènes exprimés dans les cellules érythroïdes et possédant un site pour un TF spécifique de l'érythroïde mène à un taux de validation de 50%. Une autre méthode consiste à chercher des mots surreprésentés dans un ensemble d'apprentissage de CRMs connus puis à restreindre les prédictions aux régions conservées ([Kantorovitz et al., 2009](#)). Les prédictions réalisées ont toutes été validées chez la Drosophile (5/5) comme chez la souris (2/2).

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

- **Prédiction des promoteurs**

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

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

- **Prédiction des enhancers**

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

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

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

#### 1.6.4 Validation expérimentale

Une méthode 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-

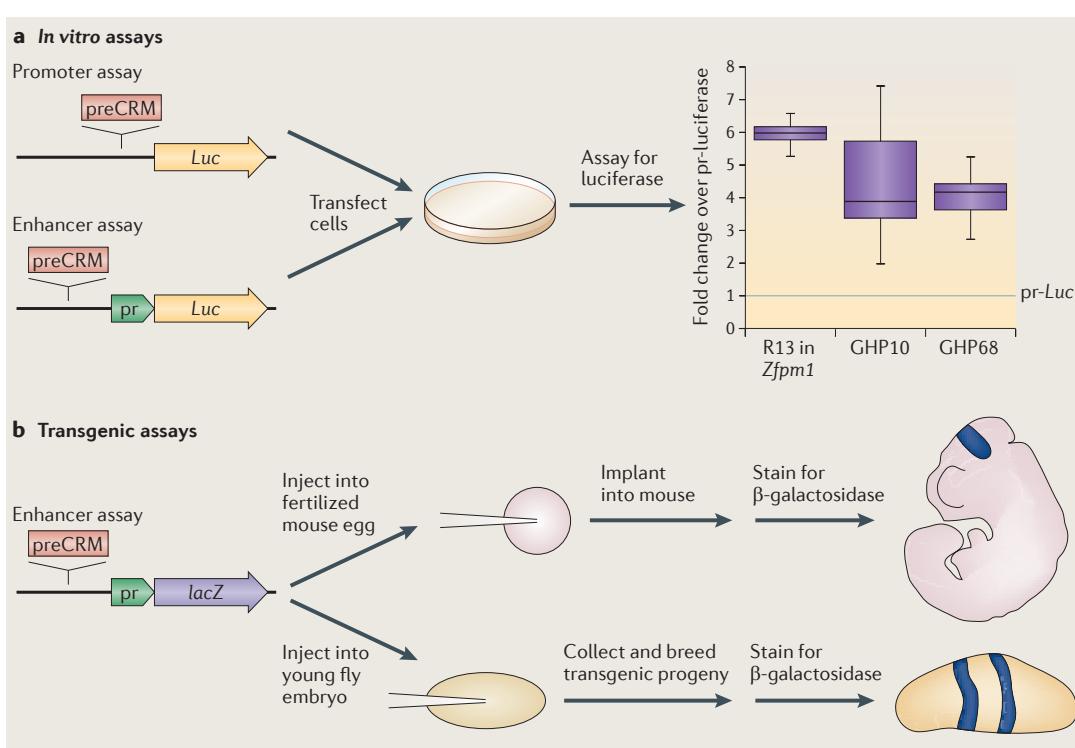


FIGURE 1.23 – Méthodes de validation des CRMs par transfection et transgenèse.

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

temporelle (tissu et stade de développement) de l'élément de régulation (fig. 1.23). Ce type d'expérience montre que le CRM prédict est *suffisant* pour reproduire le motif génétique observé. De manière optimale, il faudrait aussi montrer par délétion ciblée de l'élément de régulation au sein du génome que ce dernier est *nécessaire* à l'expression du gène endogène.



### 1.6.5 Implication des CRMs dans les maladies humaines

Au cours des dernières décennies, de nombreuses mutations dans les régions codantes des gènes, impliquant des défauts structurels des protéines associées, ont pu être associées à des maladies génétiques. À l'inverse, le rôle des mutations affectant des régions non codantes n'a été que peu exploré, essentiellement du fait de la difficulté d'annoter ces régions correctement afin de définir celles qui pourraient avoir une fonction d'intérêt. Plusieurs études ont cependant pu montrer que des variations affectant des enhancers distaux pouvaient conduire à des pathologies ([Visel et al., 2009b](#)).

L'une de ces études concerne l'enhancer spécifique du membre de *Shh* (fig. 1.24). Cet enhancer, initialement décrit chez la souris, se situe à environ 1 Mb de distance de *Shh*, au sein

*Chapitre 1. Introduction générale.*

---

de l'intron d'un gène voisin. Le séquençage de cet enhancer chez plusieurs individus humains a permis d'associer une douzaine de variations mono-nucléotidiques à la polydactylie pré axiale, c'est-à-dire la présence de doigts ou d'orteils supplémentaires (Lettice et al., 2003). Des études supplémentaires chez la souris ont montré que les variations de séquences observées dans cet enhancer conduisent à une expression ectopique dans la partie antérieure du membre au cours du développement, ce qui est consistant avec la présence de doigts supplémentaires (Masuya et al., 2007). Par ailleurs, la délétion de l'enhancer orthologue de la souris entraîne la troncation des membres (Sagai et al., 2005).

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

## 1.7 Bases de données

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

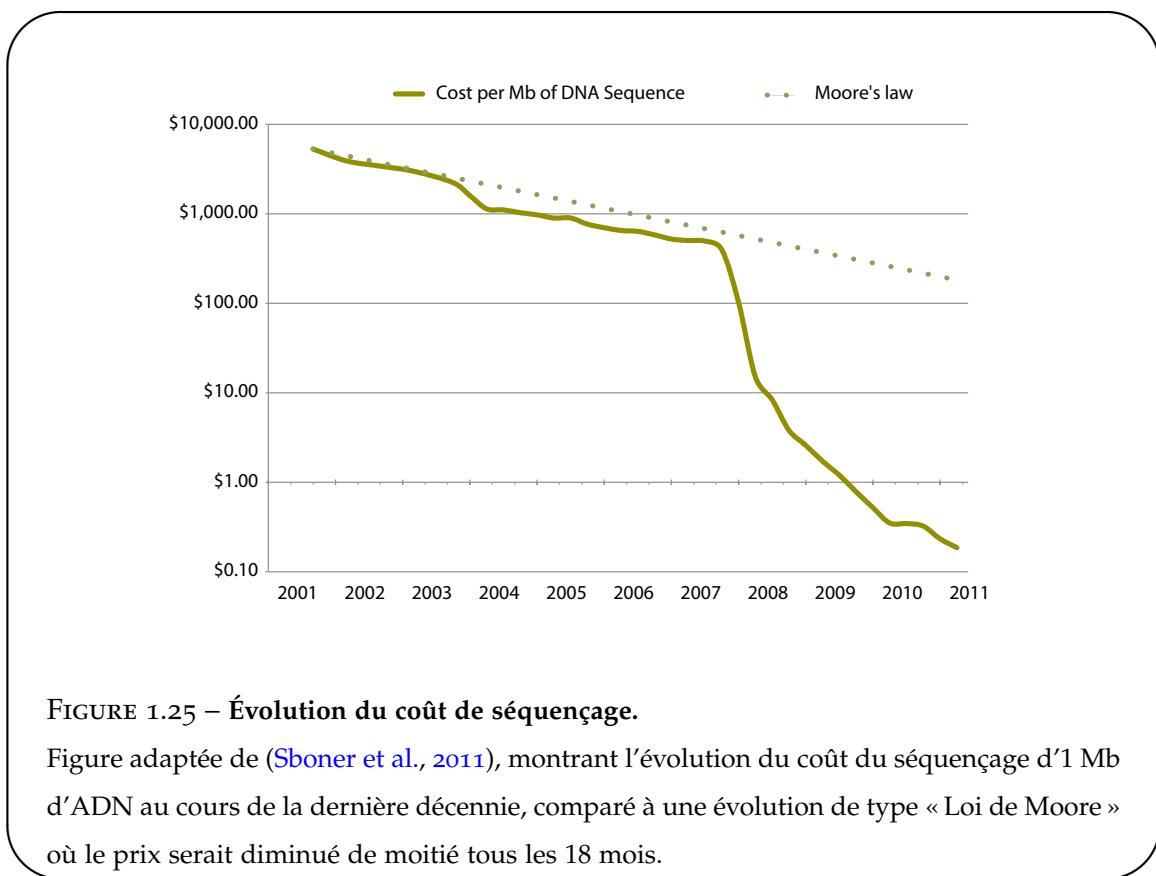


FIGURE 1.25 – Évolution du coût de séquençage.

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

### 1.7.1 Obtention de données génomiques

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

## Chapitre 1. Introduction générale.

---

et celle de l'EMBL (Ensembl, <http://www.ensembl.org>)<sup>4</sup>.

Sont à disposition les génomes des différentes espèces séquencées pour les différents assemblages réalisés, des alignements des génomes de différentes espèces deux par deux (*pair-wise alignments*) ou par groupes d'espèces (*multiple alignments*), ainsi qu'un certain nombre d'annotations essentielles à l'analyse de ces génomes : coordonnées des gènes (TSSs, exons, introns avec potentiellement différents transcrits alternatifs), miRNA ou lincRNA, ontologies associées, coordonnées des séquences répétitives (les *repeats*, en partie liés aux éléments transposables abordés en 1.5.3, et qui sont abondants dans les génomes vertébrés), différentes données ChIP-seq, indices de conservation<sup>5</sup>...

Au final, ces différentes données constituent une base de travail fiable et régulièrement mise à jour. Afin de faciliter leur obtention, il est possible d'utiliser le navigateur de tables de UCSC<sup>6</sup> ou la section BioMart d'Ensembl<sup>7</sup>.

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

En guise d'exemple, nous montrons en annexe A des statistiques obtenues aisément à partir d'annotations génétiques présentes sur UCSC et traitées avec Galaxy. Ces statistiques sont les distribution de tailles des régions intergéniques et introniques chez plusieurs espèces : la bactérie *Escherichia coli*, la levure *Saccharomyces cerevisiae*, le ver *Caenorhabditis elegans*, la mouche *Drosophila melanogaster*, la souris, le poulet et l'homme (fig. A.1).

---

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

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

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

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

### 1.7.2 Obtention de données sur les TFs

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

De nombreux modèles de TFs ont déjà été bâties préalablement à l'avènement des données haut-débit de type ChIP-seq, par exemple avec des données SELEX, et il existe des bases de données stockant les PWMs correspondantes : JAPSAR, base de donnée publique<sup>8</sup>, et TRANSFAC, qui marche par abonnement<sup>9</sup>. Il est à noter que ces PWMs ayant souvent été construites à partir d'un faible nombre de sites de fixations et de données *in vitro*, elles peuvent être relativement inadaptées à l'analyse de données *in vivo*.

### 1.7.3 Outils de visualisation

Afin d'avoir une idée plus claire des événements de régulation qui se déroulent à un locus donné, il existe plusieurs outils de visualisation des annotations génomiques et épigénétiques, que ce soit sur le site du NCBI (<http://www.ncbi.nlm.nih.gov/gene>), sur Ensembl ou sur UCSC Genome Browser. Ce dernier possède notamment l'avantage qu'il est possible d'importer des données personnelles sous un grand nombre de formats, obtenues à partir de la littérature ou à partir de ses propres travaux. Ainsi, nous présentons en figure 1.26 quelques données de ChIP-seq pour des TFs musculaires et pour des marques épigénétiques, ainsi que des prédictions bioinformatiques de sites de fixation conservés pour les homéoprotéines Six réalisée par nos soins. La visualisation sur UCSC Genome Browser permet de rapidement déterminer le mode de régulation putatif du gène *Chrng* : fixation de Six et MyoD au niveau du promoteur et apparition de marques épigénétiques H3K4me1 et H3Ac sur les histones au cours de la différenciation de progéniteurs musculaires.

Par ailleurs, il existe un outil de visualisation complémentaire de ceux cités : le visualiseur de régions conservées au cours de l'évolution ECR Browser (<http://ecrbrowser.dcode.org>), intégrant de nombreux outils bioinformatiques (Loots and Ovcharenko, 2005). Ce navigateur permet de visualiser la conservation génomique d'un locus donné chez plusieurs espèces plus ou moins lointaines (par exemple souris, homme, vache, grenouille et poisson

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

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

## Chapitre 1. Introduction générale.

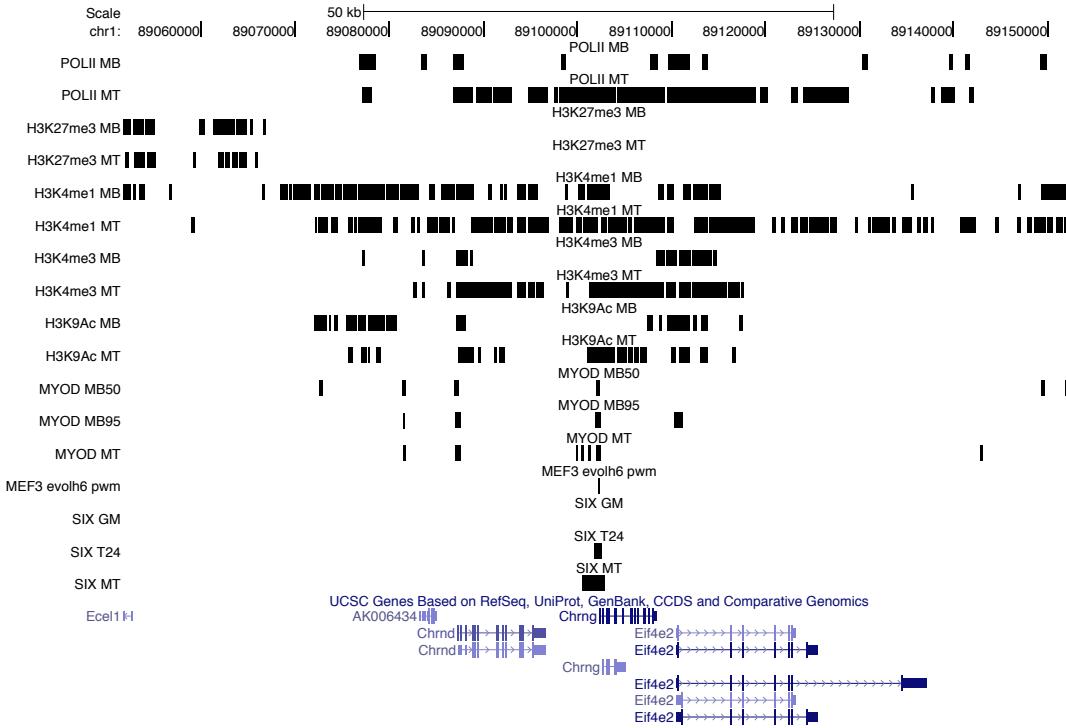
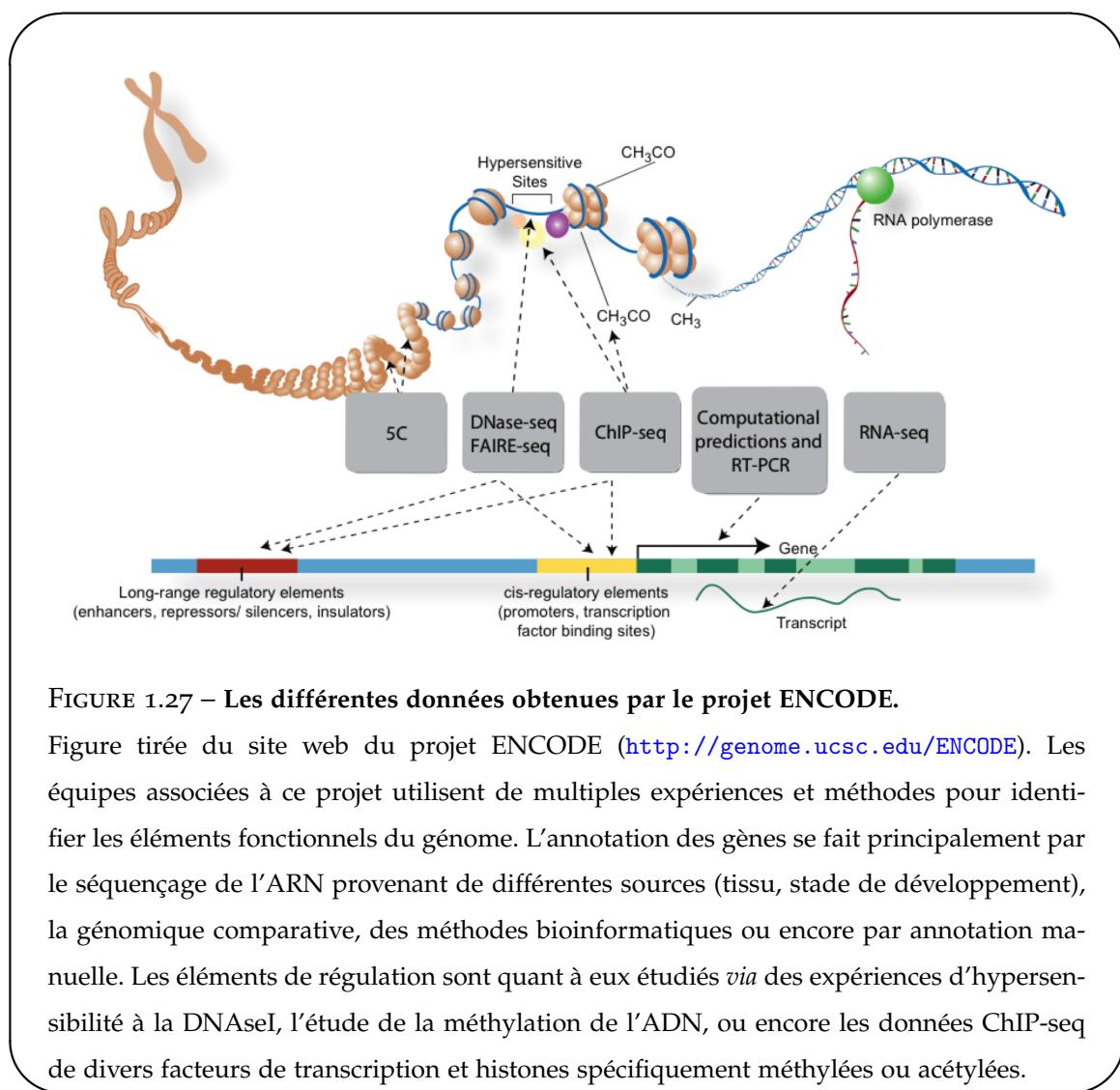


FIGURE 1.26 – Visualisation de données ChIP-seq via le site UCSC.

La visualisation de différentes données ChIP-seq et bioinformatiques (bandes noires) permet de mettre en perspective le cas de la régulation de *Chrng* (en bleu au bas de l'image) lors de la différenciation musculaire. Les données ChIP-seq sont issues de la littérature et les données bioinformatiques (MEF3) ont été obtenues au cours de cette thèse. Les rectangles noirs correspondent aux coordonnées des pics obtenus après avoir appliqué un seuil de filtrage du bruit. Les données de fixation de PolII ainsi que les données de méthylation et d'acétylation des histones (H3K4me3 et H3K9Ac, marques de l'activité transcriptionnelle, voir fig. 1.14), tirées de Asp et al. (2011), indiquent que le locus est transcrit lors de la formation de myotubes. Le TF MyoD est fixé au niveau du promoteur de *Chrng* au cours de la prolifération des myoblastes (MB50, 50% de confluence, et MB95, 95% de confluence) et au cours de la différenciation en myotubes MT (Cao et al., 2010b). Le TF Six est co-fixé avec MyoD lors de la différenciation : à T24, 24h après différenciation, et à MT (Liu et al., 2010). De plus, des analyses bioinformatiques montre l'existence dans cette région d'un site de fixation MEF3 pour la protéine Six conservé chez les vertébrés, corroborant une liaison directe de l'ADN par Six. Prises ensemble, la simple visualisation de ces données suggèrent une régulation par Six et MyoD de *Chrng*.

zèbre) afin de cibler l'étude de la régulation sur des régions extrêmement conservées. Il est ensuite possible d'analyser les séquences ultraconservées sélectionnées en utilisant les motifs de la base de donnée TRANSFAC *via* l'outil rVISTA (Loots and Ovcharenko, 2004). Un exemple d'utilisation de cet outil est donné par la découverte de plusieurs régions de régulation fonctionnelles de l'homéoprotéine Six1 possédant une extrême conservation (Sato et al., 2012).

#### 1.7.4 Le projet ENCODE



**FIGURE 1.27 – Les différentes données obtenues par le projet ENCODE.**

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

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

## Chapitre 1. Introduction générale.

---

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

Totalisant en septembre 2012 plus de 1600 expériences dans plus de 147 types cellulaires, les premières conclusions pointent vers une profusion d'événements de régulation, loin de l'idée d'ADN poubelle (*junk DNA*) : ainsi, 80% du génome est associé à un événement biochimique associé à de la formation d'ARN ou au remodelage de la chromatine, ~ 400,000 régions possèdent un état chromatinien caractéristique des enhancers et ~ 70,000 des promoteurs ([ENCODE Project Consortium et al., 2012](#)). Depuis mai 2013, les données ChIP-seq de 161 TFs couvrant 91 types cellulaires ont été mises à disposition sur UCSC Genome Browser <sup>10</sup>.

---

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

---

## Chapitre 2

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

---

<b>2.1</b>	<b>Observations de corrélations au sein des TFBS</b>	66
<b>2.2</b>	<b>Modèles existants permettant de décrire la statistique des TFBS</b>	67
<b>2.2.1</b>	Modèle de référence sans corrélations : la PWM	67
<b>2.2.2</b>	Une PWM généralisée : le modèle GWM	69
<b>2.2.3</b>	Réseaux bayésiens	70
<b>2.2.4</b>	Modèles de mélange	71
<b>2.3</b>	<b>Modèles de maximum d'entropie</b>	72
<b>2.3.1</b>	Pourquoi maximiser l'entropie ?	72
<b>2.3.2</b>	Maximisation de l'entropie sous contraintes	74
<b>2.3.3</b>	Application aux sites de fixation	75
<b>2.4</b>	<b>Article</b>	77
<b>2.5</b>	<b>Analyse thermodynamique des modèles</b>	77
<b>2.5.1</b>	Chaleur spécifique	77
<b>2.5.2</b>	Lien avec les valeurs des champs et des couplages	80
<b>2.6</b>	<b>Conclusion et perspectives</b>	80

## Introduction du chapitre 2

Dans cette partie, nous nous intéressons à la description de l'interaction entre les facteurs de transcription et leurs sites de reconnaissance sur l'ADN. Pendant longtemps, la qualité de cette description a été limitée par la quantité de données disponibles. Ainsi, les expériences de type SELEX (voir 1.4.1), où des expériences de ChIP au cas par cas permettaient de récupérer de l'ordre de quelques dizaines de sites de fixation pour un TF d'intérêt. Or, le modèle PWM, qui est le modèle le plus simple (en terme de nombre de paramètres) que l'on puisse bâtir pour décrire l'interaction possède déjà plusieurs dizaines de paramètres – les fréquences des nucléotides à chaque position –.

Ces données ne permettaient donc pas d'explorer plus en avant des modèles plus complexes de fixation incluant par exemple des termes d'interaction entre nucléotides au sein des sites de fixation. Cependant, les avancées récentes en séquençage à haut débit ont permis l'obtention de données très grande échelle, que ce soit *in vivo* par ChIP-seq ou *in vitro* par HT-SELEX (voir 1.4). Le nombre de sites de fixation obtenus est de l'ordre de quelques milliers, ce qui permet de contraindre des modèles de fixation plus complexe que le modèle PWM.

En utilisant des données ChIP-seq pour un grand nombre de facteurs de transcription de la Drosophile et des vertébrés, nous avons contraint différents modèles de fixation incluant implicitement ou explicitement des interactions entre nucléotides. Nous les avons comparés sur leur capacité à décrire les statistiques de fixation TF-ADN observées *in vivo*. Nous présentons préalablement un survol des observations et modèles existant au sujet des corrélations dans les sites de fixations de facteurs de transcription.

### 2.1 Observations de corrélations au sein des TFBS

Différents travaux ont mis en exergue l'existence de corrélations entre nucléotides au sein des sites de fixation de TFs. Parce que limitées par la quantité de données alors possible d'obtenir, les premières études de ce genre ont centré leur attention sur quelques corrélations importantes pour des cas particuliers. Ainsi, [Man and Stormo \(2001\)](#) ont observé que la protéine Mnt induit des corrélations entre les positions 16 et 17 de ses sites de reconnaissance *in vitro*. Ils ont mesuré expérimentalement la spécificité aux sites de liaisons contenant tous les variants possibles à ces deux positions. Ils ont ainsi observé que la mutation de la base consensus C en position 17 induisait un changement de préférence en position 16 de la base

---

 2.2. *Modèles existants permettant de décrire la statistique des TFBS*


---

A vers la base C. Par ailleurs, [Bulyk et al. \(2002\)](#) ont montré que la protéine EGR1 induisait des corrélations au sein d'un triplet de nucléotides central de leur site de reconnaissance. La prise en compte de ces corrélations dans l'énergie de fixation permettait alors d'améliorer la description des données par rapport au modèle additif PWM.

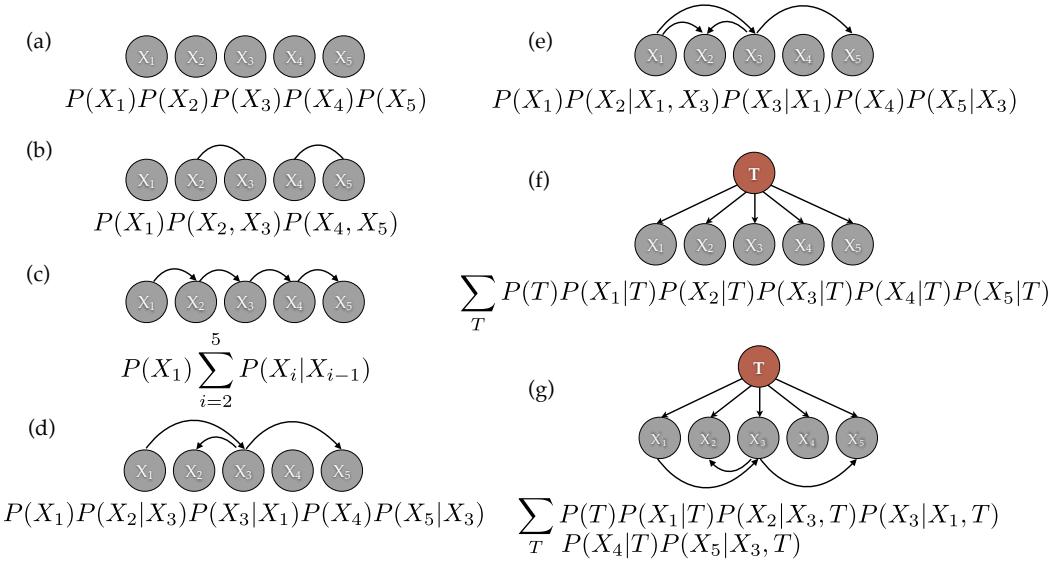
À une plus grande échelle, [Badis et al. \(2009\)](#) ont utilisé des puces à ADN (technique PBM, cf [1.4.1](#)) pour étudier la fixation *in vitro* de 104 TFs de la souris sur toutes les séquences d'ADN de 10 bp possibles. Pour chaque facteur, plusieurs centaines de séquences de fixation ont ainsi été obtenues. L'étude a révélé l'existence d'une multiplicité de motifs (PWMS) pour la plupart des TFs (seulement 15 étant mieux décrit par un motif unique). Certains motifs reconnaissent notamment des séquences à espacement variable pour lesquelles deux régions spécifiques du site sont séparées par un nombre variable de nucléotides. Enfin, les auteurs ont noté la présence de corrélations fortes dans 19 cas, celles-ci n'étant pas forcément limitées à des dinucléotides mais pouvant impliquer des trinucléotides. Plus récemment, [Jolma et al. \(2013\)](#) ont analysé par HT-SELEX plusieurs centaines de domaines de fixations à l'ADN de TFs humains et de la souris, révélant aussi l'importance d'espacements variables et surtout des corrélations dinucléotidiques entre plus proches voisins.

## 2.2 Modèles existants permettant de décrire la statistique des TFBS

Différents modèles ont été proposés pour décrire ces corrélations (fig. [2.1](#)). La méthode la plus directe consiste à partir du modèle PWM (fig. [2.1a](#)) et à ajouter des corrélations mutuellement exclusives aux positions les plus corrélées (fig. [2.1b](#)). D'autres méthodes utilisent des structures probabilistes de dépendances sous forme de chaînes de Markov (fig. [2.1c](#)) ou plus généralement de réseau bayésien ou (fig. [2.1d-e](#)). Enfin, une dernière méthode consiste à réaliser des mélanges de modèles afin de capturer des ensembles distincts de corrélations (fig. [2.1f-g](#)).

### 2.2.1 Modèle de référence sans corrélations : la PWM

Nous l'avons vu, le modèle le plus simple (en termes de nombre de paramètres) décrivant l'interaction entre un TF et son site de reconnaissance sur l'ADN consiste à faire l'hypothèse que les nucléotides contribuent indépendamment à l'énergie de fixation. Cette hypothèse



**FIGURE 2.1 – Différents modèles pour décrire les corrélations entre nucléotides dans les sites de fixation de facteurs de transcription.**

Figure adaptée de Barash et al. (2003) illustrant différents modèles de fixation sur un site de longueur 5. Pour chaque modèle, la structure du réseau de dépendances sous-jacent est représentée, ainsi que la distribution de probabilité  $P(X_1, X_2, X_3, X_4, X_5)$  correspondante, où  $X_i$  est une variable aléatoire prenant la valeur (A, C, G ou T) du nucléotide à la position  $i$ . Les modèles représentés sont les suivants : (a) PWM (pas de corrélations), (b) GWM (corrélations mutuellement exclusives), (c) chaîne de Markov d'ordre 1 (corrélations entre plus proches voisins), (d) réseau bayésien en arbre (au plus un parent par noeud) ou (e) pas en arbre (le noeud 2 a deux parents), (f) mélange de PWMs et (e) mélange d'arbres à dépendances fixées.

conduit au modèle PWM (section 1.3.2 et fig.2.1a), qui s'écrit<sup>11</sup> :

$$P(X_1, \dots, X_k) = \prod_{i=1}^K P(X_i) \quad (2.1)$$

où  $P(X_i)$  est la probabilité marginale d'observer le nucléotide  $X \in \{A, C, G, T\}$  à la position  $i$ . Un tel modèle possède  $3K$  paramètres – 3 paramètres  $P(X_i)$  par position, la normalisation des probabilités permettant de fixer le paramètre restant –. Pour une longueur de site typique  $K = 10$ , le modèle PWM contient 30 paramètres à contraindre, sachant qu'un « modèle »

<sup>11</sup>. Comme nous l'avons signalé en 1.3.2, le terme PWM (*Position Weight Matrix*) réfère en fait à la matrice des poids  $\log(P(X_i)/\pi_{X_i})$  où  $\pi_{X_i}$  est une distribution neutre indépendante de la position (dite distribution *background*), par exemple calculée sur des régions intergéniques.

complet paramétrant la distribution jointe sans faire d'hypothèse comporterait  $4^{10} - 1 \sim 10^6$  paramètres.

### 2.2.2 Une PWM généralisée : le modèle GWM

Une première méthode permettant de complexifier le modèle PWM consiste à intégrer explicitement des groupes mutuellement exclusif<sup>12</sup> de nucléotides corrélés au sein du modèle (fig. 2.1b). Une telle méthode fut d'abord employée par Benos et al. (2002) pour prendre en compte des corrélations préalablement définies entre nucléotides plus proches voisins. De manière plus générale, Zhou and Liu (2004) ont développé un modèle de matrice de poids généralisée (GWM pour *Generalized Weight Matrix*) qui prend en compte de manière systématique les corrélations permettant d'améliorer le modèle indépendant. Pour ce faire, les auteurs utilisent une méthode de Monte-Carlo par chaîne de Markov (MCMC) : des corrélations sont ajoutées ou enlevées au hasard au modèle et acceptées selon la règle de Metropolis-Hastings (Krauth, 2006). Cette acceptation est proportionnelle au facteur de Bayes, une quantité qui permet de comparer des modèles possédant des nombres de paramètres différent<sup>13</sup>. Ce facteur est défini par le rapport entre la probabilité de générer les données  $D$  (les séquences de fixation) avec un modèle  $M_1$  de paramètres  $\theta_1$  plutôt qu'avec un autre modèle  $M_2$  de paramètres  $\theta_2$  :

$$BF = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(D|\theta_1, M_1)P(\theta_1|M_1)d\theta_1}{\int P(D|\theta_2, M_2)P(\theta_2|M_2)d\theta_2} \quad (2.2)$$

Le modèle final consiste en un ensemble de paramètres décrivant des positions indépendantes et des positions corrélées mutuellement exclusives. En analysant les données TRANSFAC, les auteurs ont noté que dans 25% des cas (22/95) le modèle GWM était significativement meilleur que le modèle PWM (facteur de Bayes supérieur à 6).

Cette méthode a par la suite été utilisée sur des données ChIP-seq pour 4 TFs mammifères – NRSF, STAT1, CTCF et ER – (Hu et al., 2010). En utilisant les 10% des pics les plus importants comme ensemble d'apprentissage et en se restreignant aux régions de 200bp centrées autour du sommet du pic ChIP, les auteurs ont réalisé un échantillonnage de Gibbs (Casella and George, 1992) pour obtenir les sites de fixation suivant les hypothèses que (1) chaque pic contient au plus un seul site de fixation (modèle ZOOPS pour *Zero or One Occurrences Per*

12. Les corrélations entre des couples de positions (i,j) et (j,k) ne peuvent être admises au sein du même modèle.

13. Sous certaines approximation, ce facteur peut se rapporter à une différence de valeurs du BIC, introduit dans l'article en 2.4.

Sequence), (2) la probabilité *a priori* d'avoir un site à une certaine position sur la séquence est plus forte autour du sommet du pic, et (3) les sites sont décrits par un modèle GWM. L'étude a révélé l'existence de corrélations fortes limitées aux nucléotides plus proches voisins dans les quatre cas étudiés. Les nucléotides participant aux corrélations se situaient à des positions ayant un faible contenu en information dans le modèle PWM. Enfin, les auteurs ont noté la présence de plusieurs triplets de nucléotides voisins corrélés.

### 2.2.3 Réseaux bayésiens

Une généralisation du modèle GWM consiste à supprimer la condition d'exclusion mutuelle des paires de nucléotides corrélés en décrivant de manière plus générale le réseau de dépendance entre positions. Une telle description est possible en utilisant le langage des réseaux bayésiens. Les dépendances y sont représentées par un graphe orienté acyclique<sup>14</sup>  $G$ , dont les nœuds sont les variables  $X_i$  et les liens représentent les conditionnements d'une variable avec des variables parentes (fig. 2.1e). La probabilité jointe s'écrit :

$$P(X_1, \dots, X_k) = \prod_{i=1}^K P(X_i | P_i^G) \quad (2.3)$$

où  $P_i^G$  est l'ensemble (pouvant être vide) des parents de  $X_i$  dans  $G$ . Le nombre de paramètres peut rapidement devenir grand : si l'on note  $N_i$  le nombre de parents de  $X_i$ , alors le nombre de paramètres du modèle est  $3 \sum_{i=1}^K 4^{N_i}$ .

Lorsque les différents nœuds possèdent au plus un parent, on parle d'arbre bayésien (fig. 2.1d). Ce type de réseau bayésien généralise notamment le cas des chaînes de Markov d'ordre 1, où chaque nœud dépend du nœud précédent (fig. 2.1c). Le nombre de paramètres est alors restreint, puisqu'il est au plus de  $3 \cdot 4K$ .

L'avantage des arbres bayésiens est qu'il existe des algorithmes efficaces permettant de trouver la meilleure structure d'arbre (Friedman et al., 1997). De tels modèles d'arbres ont été utilisés pour décrire les données de 95 TFs de Transfac (Barash et al., 2003). Dans  $\sim 25\%$  des cas (22/95), le modèle d'arbre bayésien s'avère significativement meilleur qu'un modèle PWM, ce qui est du même ordre de grandeur que pour le modèle GWM vu en 2.2.2.

<sup>14</sup>. Un graphe orienté acyclique est un réseau dont les liens sont orientés et au sein duquel il n'est pas possible de revenir à son point de départ en suivant les flèches

### 2.2.4 Modèles de mélange

Dans les cas précédents, nous avons présenté des modèles capturant des dépendances « locales » entre paires de nucléotides. Néanmoins, il peut exister des dépendances plus largement réparties entre les positions, comme cela a déjà été observé empiriquement ([Badis et al., 2009](#); [Jolma et al., 2013](#)). De telles corrélations à plus grande échelle peuvent être modélisées en supposant que le facteur de transcription possède plusieurs « modes » de fixation. Ceux-ci peuvent par exemple correspondre à différentes conformations de la protéine sur son site de fixation, chaque configuration possédant ses propres préférences de fixation. Ces modes sont décrits par une variable aléatoire  $T$  (le *type* de fixation) de probabilité  $P(T)$ . Il est ensuite possible de décrire la fixation au sein de chaque mode par l'un des modèles précédents.

- **Mélange de PWMs**

Le cas le plus naturel consiste à utiliser comme modèle de fixation le modèle PWM, c'est-à-dire que dans chaque mode il y a indépendance entre les positions. La probabilité d'observer un site est alors donnée par la somme sur les différents modes de fixation de la probabilité de fixer un site, conditionnée par la probabilité d'être dans ce mode :

$$P(X_1, \dots, X_K) = \sum_{T=1}^N P(T) \prod_{i=1}^K P(X_i|T) \quad (2.4)$$

où  $N$  est le nombre de modes de fixation. Ce modèle a plusieurs avantages. D'abord, le nombre de paramètres reste linéaire en  $K$  : pour décrire  $P(T)$  et les  $N$  PWMs il faut  $N - 1 + 3KN$  paramètres. Ce nombre reste donc raisonnablement faible devant le nombre de paramètres requis pour complètement décrire les interactions à deux nucléotides, qui croît comme  $K^2$ . Ensuite, le modèle a une interprétation claire qui peut permettre de mettre en exergue un mécanisme biologique sous-jacent.

Ce type de modèle permet de dépasser le modèle PWM dans un nombre substantiel de cas. Ainsi, [Barash et al. \(2003\)](#) ont montré que  $\sim 40\%$  des TFs de Transfac (36/95) sont significativement mieux représentés par un mélange de 2 PWMs que par une seule PWM. En utilisant des données *in vitro* plus précises pour 104 TFs de la souris, [Badis et al. \(2009\)](#) ont montré que  $\sim 85\%$  (89/104) étaient mieux représenté par une combinaison de PWMs que par une PWM seule, plaident pour un portée générale de l'existence de « motifs secondaires ».

- **Mélange d'arbres**

De la même manière que les PWMs, il est possible d'étendre les modèles d'arbres en réalisant un mélange d'arbres. Intuitivement, ceci permet de capturer des dépendances additionnelles en gardant un nombre de paramètres linéaire en fonction de la taille du motif. Un tel modèle semble posséder des performances comparables au mélange de PWM, et améliore la description des TFs de Transfac dans  $\sim 40\%$  des cas (35/95) ([Barash et al., 2003](#)).

## 2.3 Modèles de maximum d'entropie

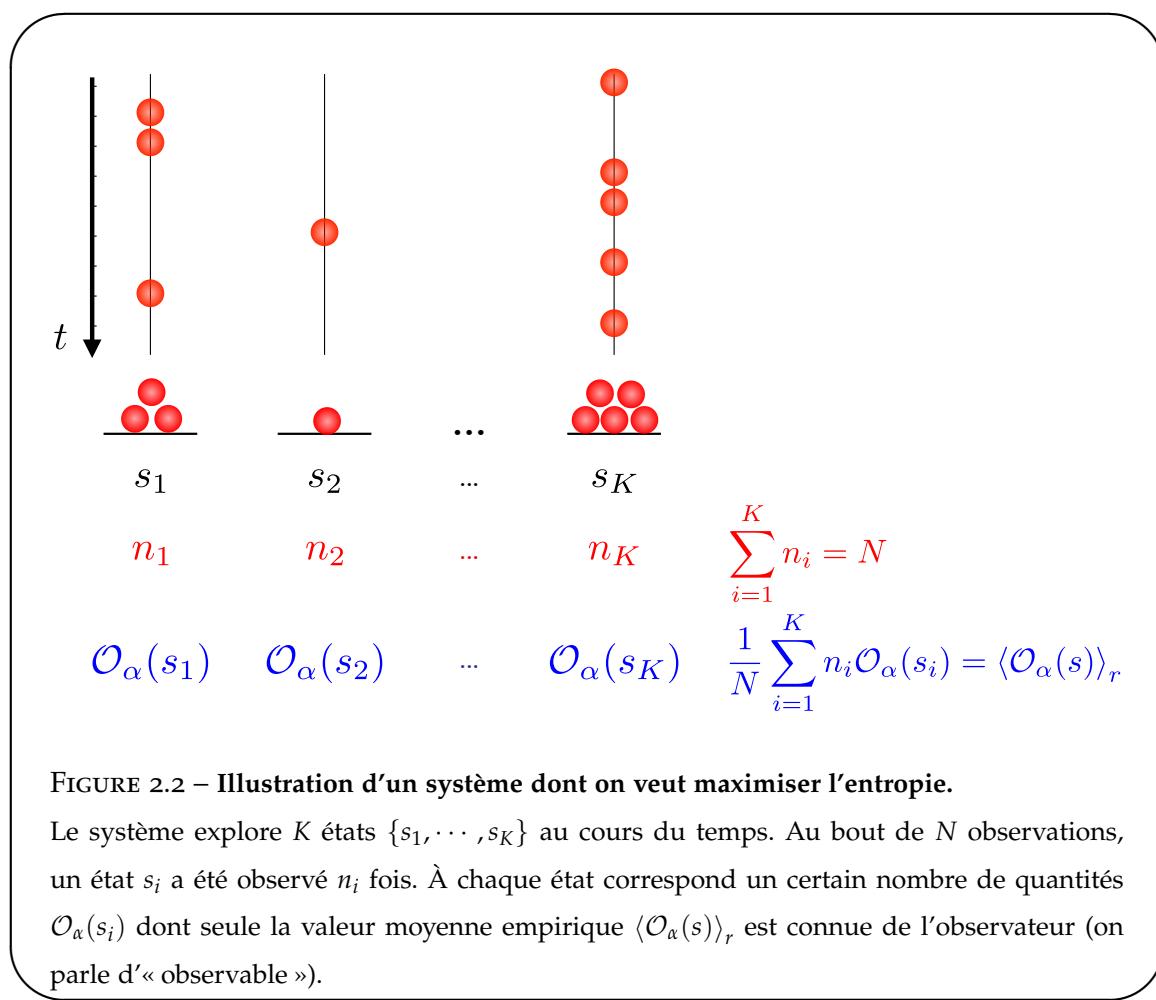


FIGURE 2.2 – Illustration d'un système dont on veut maximiser l'entropie.

Le système explore  $K$  états  $\{s_1, \dots, s_K\}$  au cours du temps. Au bout de  $N$  observations, un état  $s_i$  a été observé  $n_i$  fois. À chaque état correspond un certain nombre de quantités  $\mathcal{O}_\alpha(s_i)$  dont seule la valeur moyenne empirique  $\langle \mathcal{O}_\alpha(s) \rangle_r$  est connue de l'observateur (on parle d'« observable »).

### 2.3.1 Pourquoi maximiser l'entropie ?

Le concept d'entropie remonte aux prémisses de la physique statistique ([Jaynes, 1978](#)). Dans l'essence, il peut être compris de la manière suivante. Supposons qu'un système com-

porte  $K$  états distincts  $\{s_1, \dots, s_K\}$ . Au cours du temps, le système explore les différents états (fig. 2.2). Au bout de  $N$  observations, chaque état a été observé un nombre  $n_i$  de fois. La question sous-jacente au calcul de l'entropie est la suivante : sans connaissance *a priori* sur le système, que puis-je dire de ces  $n_i$ ? Prenons l'exemple de la figure 2.2. On a certaines valeurs pour les  $n_i$  ( $n_1 = 3$ ,  $n_2 = 1$ , etc.), et on aimerait savoir de combien de manières il est possible de réaliser un tel ensemble de valeurs. Notons ce nombre  $\mathcal{N}(n_1, \dots, n_K)$ . Il est donné par la formule suivante :

$$\begin{aligned}\mathcal{N}(n_1, \dots, n_K) &= \binom{N}{n_1} \binom{N-n_1}{n_2} \cdots \binom{N-\sum_{i=1}^{K-1} n_i}{n_K} \\ &= \frac{N!}{(N-n_1)!n_1!} \times \frac{(N-n_1)!}{(N-n_1-n_2)!n_2!} \times \cdots \times \frac{(N-\sum_{i=1}^{K-1} n_i)!}{0!n_K!}\end{aligned}\quad (2.5)$$

soit

$$\mathcal{N}(n_1, \dots, n_K) = \frac{N!}{n_1!n_2!\cdots n_K!} \quad (2.6)$$

Il convient alors de s'intéresser au logarithme de cette quantité. En effet, dans le cas où les nombres d'observation sont grands  $n_i \gg 1$ , ceux-ci s'expriment simplement grâce à la formule de Stirling :

$$\log(n!) \xrightarrow{n \rightarrow \infty} n \log(n) - n \quad (2.7)$$

On peut alors écrire

$$\begin{aligned}\log \mathcal{N}(n_1, \dots, n_K) &= N \log(N) - N - \sum_{i=1}^K (n_i \log(n_i) - n_i) \\ &= \sum_{i=1}^K n_i \log\left(\frac{N}{n_i}\right) \\ &= -N \sum_{i=1}^K \frac{n_i}{N} \log\left(\frac{n_i}{N}\right)\end{aligned}\quad (2.8)$$

On note l'apparition des probabilités empiriques  $f(s_i) = n_i/N$  d'observer l'état  $s_i$ , qui tendent asymptotiquement (dans la limite « thermodynamique »  $N \rightarrow \infty$ ) vers les « vraies » probabilités  $P(s_i)$ . L'entropie est définie dans cette limite comme étant égale à  $1/N \log \mathcal{N}(n_1, \dots, n_K)$ , soit

$$S[P] = - \sum_{\{s\}} P(s) \log P(s) \quad (2.9)$$

où  $\{s\} = \{s_1, \dots, s_K\}$  dénote l'ensemble des états accessibles. L'idée est alors la suivante : nous souhaitons savoir quels états le système a le plus probablement visité au cours des  $N$  transitions. Sans connaissance *a priori* sur le système, il est plus probable que les nombres  $(n_1, \dots, n_K)$  obtenus soient ceux qui sont réalisés le plus souvent, c'est-à-dire ceux qui maximisent la quantité  $\mathcal{N}(n_1, \dots, n_K)$  et donc au final l'entropie. Par ailleurs, les fluctuations relatives des quantités  $n_i$  sont de l'ordre de  $1/\sqrt{n_i}$  (Sethna, 2006). Ainsi, la solution de maximum d'entropie domine largement les autres solutions possibles dans la limite thermodynamique.

### 2.3.2 Maximisation de l'entropie sous contraintes

Notons  $\mathcal{O}_\alpha(s)$  une quantité attachée à  $s$  (fig. 2.2). En thermodynamique, une telle quantité correspond par exemple à l'énergie d'un état. L'observateur n'a lui accès qu'aux valeurs moyennes de telles quantités sous-jacentes. À l'état d'équilibre thermodynamique, l'échantillonnage des états est réalisé au sein de la distribution de probabilité de maximum d'entropie  $P(s)$ , et les valeurs moyennes calculées avec les fréquences empiriques  $f(s)$  doivent donc être compatibles avec les valeurs moyennes calculées avec la distribution de probabilité  $P(s)$  :

$$\sum_{\{s\}} P(s) \mathcal{O}_\alpha(s) = \sum_{\{s\}} f(s) \mathcal{O}_\alpha(s) \quad (2.10)$$

Nous souhaiterions maintenant connaître la distribution  $P(s)$  la moins biaisée (i.e de maximum d'entropie) qui satisfait les contraintes de l'éq. 2.10 imposées par l'observation des données (l'information que possède l'observateur). Ce problème revient à maximiser le Lagrangien suivant :

$$\mathcal{L} = - \sum_{\{s\}} P(s) \log P(s) + \lambda \left( \sum_{\{s\}} P(s) - 1 \right) + \sum_\alpha \beta_\alpha \sum_{\{s\}} (P(s) - f(s)) \mathcal{O}_\alpha(s) \quad (2.11)$$

où les paramètres  $\lambda$  et  $\beta_\alpha$  sont les multiplicateurs de Lagrange correspondant respectivement à la contrainte de normalisation de la distribution de probabilité et aux informations qu'a l'observateur sur certaines valeurs moyennes du système (éq. 2.10). La maximisation de ce Lagrangien est obtenue en annulant la dérivée fonctionnelle par rapport à la distribution de probabilité  $P(s)$  :

$$\frac{\delta \mathcal{L}}{\delta P(s)} = 0 = -\ln P(s) - 1 + \lambda + \sum_{\alpha} \beta_{\alpha} \mathcal{O}_{\alpha}(s) \quad (2.12)$$

En utilisant la normalisation des probabilités, il est possible de trouver  $\lambda$ , et la solution se met finalement sous la forme

$$P(s) = \frac{1}{Z} e^{-\mathcal{H}(s)} \quad (2.13)$$

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

$$\mathcal{H} = \sum_{\alpha} \beta_{\alpha} \mathcal{O}_{\alpha}(s) \quad (2.14)$$

et  $Z$  est la fonction de partition permettant la normalisation de la distribution  $P(s)$  :

$$Z = \sum_{\{s\}} \exp[-\mathcal{H}(s)]. \quad (2.15)$$

### Remarque

Il est possible de montrer que la maximisation de l'entropie, partant des contraintes de l'éq. 2.10 sur les valeurs moyennes pour en arriver à une forme exponentielle de la distribution de probabilité, est le contrepoint d'une maximisation de la vraisemblance partant d'une forme exponentielle pour en arriver aux mêmes contraintes sur les valeurs moyennes ([Grendar Jr and Grendar, 2001](#); [Jaynes, 1982](#)).

### 2.3.3 Application aux sites de fixation

- **Corrélations à un point : le modèle PWM**

Dans le cas qui nous intéresse, un état  $s$  correspond à une séquence d'ADN appartenant à l'ensemble  $\{s\}$  des sites de fixation d'un facteur de transcription. Considérons maintenant l'observable quantifiant la présence du nucléotide  $a$  à la position  $i$  d'un site :

$$\mathcal{O}_{i,a}(s) = \delta(s_i, a) \quad (2.16)$$

où  $\delta$  est la fonction de Kronecker qui vaut 1 lorsque le nucléotide à la position  $i$  du site  $s_i$  vaut  $a$  et 0 sinon. De cette définition il suit que la valeur moyenne sur les fréquences empiriques

$$\sum_{\{s\}} f(s) \mathcal{O}_{i,a}(s) = f_{i,a} \quad (2.17)$$

se réduit à la fréquence du nucléotide  $a$  à la position  $i$ . Notons  $h_i(a)$  le multiplicateur de Lagrange correspondant et  $\mathcal{A} = \{A, C, G, T\}$ . On trouve alors

$$\begin{aligned} \mathcal{H}(s) &= \sum_{i=1}^L \sum_{a \in \mathcal{A}} h_i(a) \delta(s_i, a) \\ &= \sum_{i=1}^L h_i(s_i) \end{aligned} \quad (2.18)$$

Les différentes positions étant indépendantes, la fonction de partition  $\mathcal{Z}$  peut par ailleurs se scinder en différentes fonctions de partition par position :  $\mathcal{Z} = \prod_{i=1}^L \mathcal{Z}_i$ . On obtient au final

$$P(s) = \frac{1}{\mathcal{Z}} e^{-\sum_{i=1}^L h_i(s_i)} = \prod_{i=1}^L \frac{e^{-h_i(s_i)}}{\mathcal{Z}_i} \quad (2.19)$$

On retrouve le modèle PWM introduit dans l'éq. 2.13.

- **Corrélations à deux points : le modèle de Potts**

Il est maintenant relativement direct de complexifier le modèle en ajoutant l'observation des couples d'interaction au sein des sites de fixation :

$$\mathcal{O}_{i,a,j,b}(s) = \delta(s_i, a) \delta(s_j, b) \quad (2.20)$$

La corrélation à deux points entre le nucléotide  $a$  en position  $i$  et  $b$  en position  $j$  s'écrit donc

$$\sum_{\{s\}} f(s) \mathcal{O}_{i,a,j,b}(s) = f_{i,a,j,b} \quad (2.21)$$

où  $f_{i,a,j,b}$  est la fréquence empirique d'observation de la paire de nucléotide  $(a, b)$  aux positions  $(i, j)$ . Notons  $J_{i,j}(a, b)$  le multiplicateur de Lagrange correspondant. L'Hamiltonien sous les contraintes imposées par les équations 2.17 et 2.21 s'écrit :

$$\begin{aligned} \mathcal{H}(s) &= \sum_{i=1}^L \sum_{a \in \mathcal{A}} h_i(a) \delta(s_i, a) + \sum_{i=1}^{L-1} \sum_{j>i} \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} J_{i,j}(a, b) \delta(s_i, a) \delta(s_j, b) \\ &= \sum_{i=1}^L h_i(s_i) + \sum_{i=1}^{L-1} \sum_{j>i} J_{i,j}(s_i, s_j) \end{aligned} \quad (2.22)$$

Le modèle de maximum d'entropie est finalement

$$P(s) = \frac{1}{Z} e^{-\sum_{i=1}^L h_i(s_i) - \sum_{i=1}^{L-1} \sum_{j>i} J_{i,j}(s_i, s_j)} \quad (2.23)$$

On reconnaît le modèle de Potts inhomogène de champs magnétiques locaux  $h_i$  et de termes d'interaction  $J_{i,j}$  couramment utilisé dans la description des verres de spins (Baxter, 2007).

## 2.4 Article

L'article qui suit décrit l'analyse de données de fixation *in vivo* à grande échelle pour plusieurs TFs drosophiles et mammifères. Différents modèles sont comparés, incluant ou non des dépendances : un modèle PWM, un modèle de mélange de PWMs, et un modèle de Potts.

## 2.5 Analyse thermodynamique des modèles

### 2.5.1 Chaleur spécifique

En plus des résultats présentés dans l'article, nous nous sommes intéressés à une quantité classique de la thermodynamique : la chaleur spécifique ou capacité calorifique. Considérons un modèle décrit par la statistique de Boltzmann à la température inverse  $\beta = 1/T$  (on omet la constante de Boltzmann  $k$  en l'intégrant à l'énergie) :

$$P(s) = \frac{1}{Z} e^{-\beta E(s)} \quad (2.24)$$

Le cas de l'équation 2.23 correspond au cas particulier  $\beta = 1$ . Nous voulons voir comment l'amplification ou la diminution globale de l'écart entre les énergies affecte la possibilité du système d'explorer les différents états possibles. À température nulle ( $T \rightarrow 0$  ou  $\beta \rightarrow \infty$ ), le système reste dans le niveau fondamental de minimum d'énergie et de probabilité 1, alors qu'à des températures non nulles le système à l'énergie  $E_0$  transite vers un état d'énergie supérieure  $E_1$  avec une probabilité  $\propto \exp(-\beta(E_1 - E_0))$ . Lorsqu'un paysage énergétique est composé de plusieurs puits d'énergie séparés par des barrières énergétiques importantes, on s'attend à avoir une (ou plusieurs) températures critiques à partir desquelles de fortes

différences d'énergie deviennent franchissables. L'énergie moyenne peut alors être significativement affectée, sautant soudainement à une nouvelle valeur du fait du poids des nouveaux états explorés.

La chaleur spécifique permet de caractériser ces sauts soudains d'énergie moyenne lors de la variation de la température, caractéristiques des transitions de phase. Elle mesure simplement la variation de l'énergie moyenne lors d'une variation de température :

$$C(T) = \frac{d\langle E \rangle}{dT} \quad (2.25)$$

où

$$\langle E \rangle = \sum_{\{s\}} E(s) \frac{e^{-\beta E(s)}}{\mathcal{Z}} \quad (2.26)$$

Cette chaleur spécifique peut par ailleurs s'écrire sous une forme plus utile :

$$\begin{aligned} \frac{d\langle E \rangle}{dT} &= -\beta^2 \frac{d\langle E \rangle}{d\beta} \\ &= -\beta^2 \left[ \sum_{\{s\}} E(s) \left( -E(s)e^{-\beta E(s)} \right) \frac{1}{\mathcal{Z}} + \sum_{\{s\}} E(s)e^{-\beta(E(s))} \left( -\frac{d\mathcal{Z}}{d\beta} \frac{1}{\mathcal{Z}^2} \right) \right] \\ &= \beta^2 \left[ \langle E^2 \rangle - \langle E \rangle^2 \right] \end{aligned} \quad (2.27)$$

Ainsi, la chaleur spécifique  $C(T)$  est directement accessible en regardant les corrélations de l'énergie sur l'ensemble des états du système, ce qui peut se calculer simplement à partir des modèles de fixation. Nous avons calculé la variation de  $C(T)$  en fonction de la température pour les modèles indépendant et avec dépendances obtenus en 2.4 pour les différents TFs étudiés. Une température fictive est introduite dans les modèles en multipliant les énergies par  $\beta$ , afin de se placer dans le cadre de l'équation 2.24. Les résultats sont montrés en figure 2.3 (modèle indépendant en bleu, modèle de Potts en rouge). On observe pour la plupart des facteurs l'existence de deux pics de chaleur spécifique pour des températures de l'ordre de  $T \sim 10^{-1}$  et  $T \sim 5$  (par exemple,  $T = 0.4$  et  $T = 2.8$  dans le cas du modèle indépendant de Twist). Il y a de légères variations entre les deux modèles : notamment, le premier pic semble renforcé par le modèle de Potts dans plusieurs cas (par exemple, E2f4, NRSF, TCF3 ou Twist). Néanmoins, le nombre de pics (ou de transitions de phases) reste le même.

## 2.5. Analyse thermodynamique des modèles

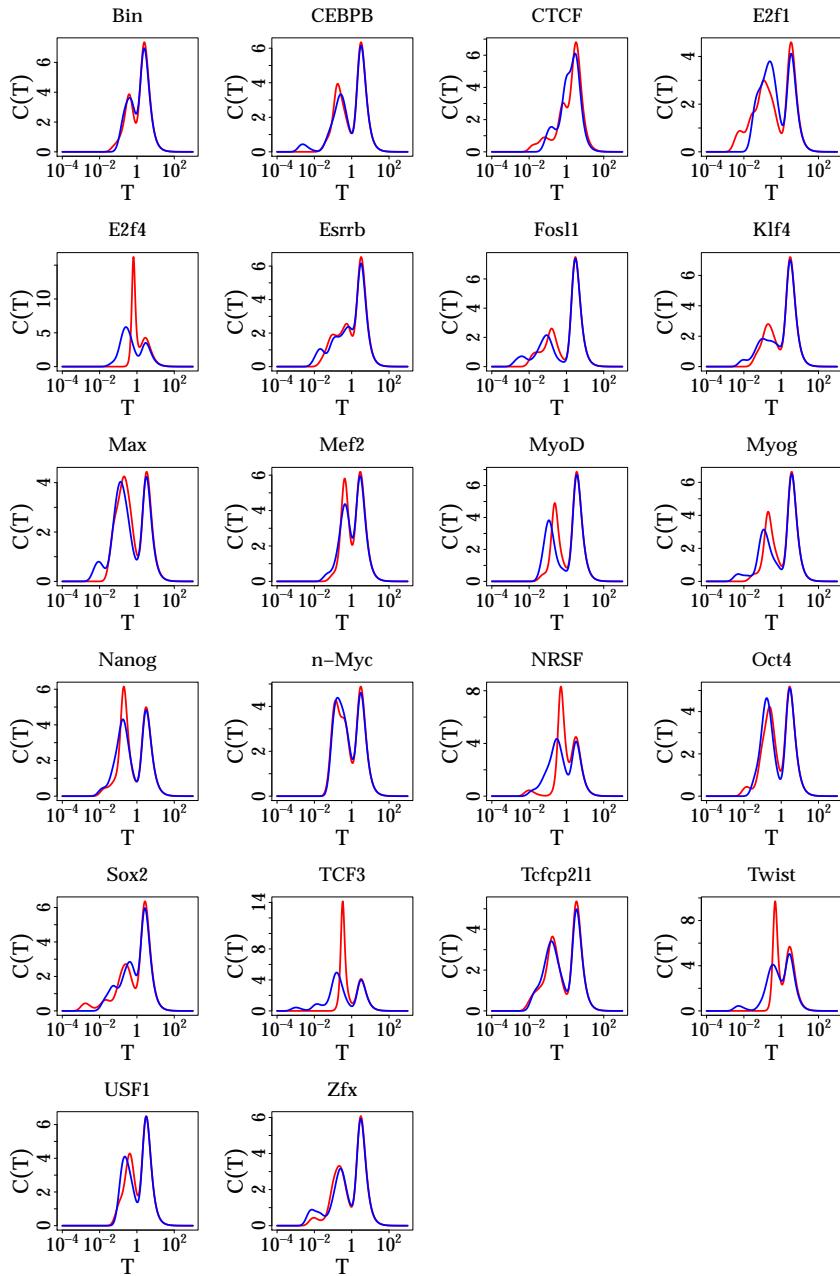


FIGURE 2.3 – Chaleur spécifique pour différents TFs.

La chaleur spécifique (l'équivalent de la capacité calorifique en thermodynamique)  $C(T) = d\langle E \rangle / dT$  est tracée en fonction de la température  $kT$  (échelle logarithmique) pour les différents TFs considérés. Le modèle indépendant (bleu) et le modèle de Potts avec interactions (rouge) sont comparables dans la plupart des cas.

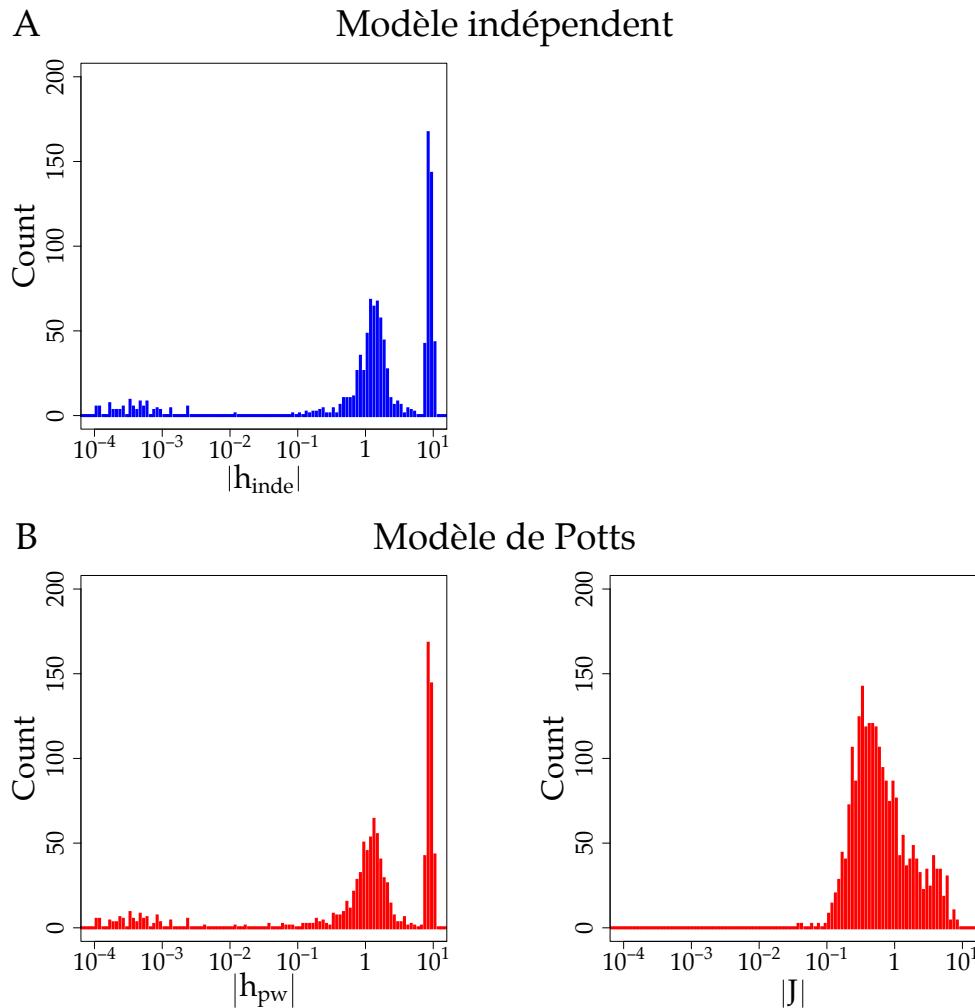
### 2.5.2 Lien avec les valeurs des champs et des couplages

Afin de comprendre l'existence des pics de chaleur spécifique et les énergies (températures) associées, il faut revenir aux modèles d'énergie. Lorsque l'on regarde l'histogramme des valeurs absolues de  $h_i$  obtenues dans les modèles indépendant des différents TFs étudiés, on trouve plusieurs valeurs typiques autour de  $10^{-4}$ , 1 et 10 (fig. 2.4A). Celles-ci peuvent s'expliquer de la manière suivante. Dans le modèle indépendant, les champs sont simplement le logarithme naturel de la probabilité d'observer un nucléotide  $a$  à une position  $i$  donnée  $h_i(a) = -\log P_i(a)$  (la jauge est choisie telle que  $\mathcal{Z}_i = 1$ ). En valeur absolue, les champs  $h_i$  proches de 0 ( $h_i \sim 10^{-4} - 10^{-3}$ ) correspondent aux nucléotides très conservés (toujours observés), les valeurs autour de 1 correspondent à des nucléotides dégénérés (i.e également observés :  $|\log(1/4)| \sim 1.4$ ) et les valeurs autour de 10 correspondent aux nucléotides qui ne sont jamais observés, au pseudocount près (pour un pseudocount de 1 et  $10^4$  séquences,  $|\log(10^{-4})| \sim 9.2$ ). On peut maintenant mieux comprendre les pics de chaleur spécifique. À température nulle, seuls les sites consensus sont accessibles. Lorsque la température se rapproche de 1, les nucléotides dégénérés d'énergie  $h_i \sim 1$  deviennent accessibles, augmentant significativement la valeur de l'énergie moyenne (premier pic). Puis, lorsque la température se rapproche de 10, les nucléotides non observés d'énergie  $h_i \sim 10$  deviennent à leur tour accessibles, augmentant à nouveau l'énergie moyenne (deuxième pic).

Dans le cas du modèle de Potts (fig. 2.4B), les champs  $h_i$  prennent des valeurs proches de celles obtenues avec le modèle indépendant. Par ailleurs, les interactions  $J_{i,j}$  sont réparties autour d'un mode centré autour de  $J_{i,j} \sim 0.5$ , ce qui correspond l'échelle d'énergie du premier pic. Ainsi, le renforcement du premier pic de chaleur spécifique par rapport au cas indépendant observé pour plusieurs TFs de la figure 2.3 peut s'expliquer par l'effet des termes d'interaction  $J_{i,j}$ .

## 2.6 Conclusion et perspectives

Nous avons analysé les dépendances au sein des sites de fixation liés *in vivo* pour différents facteurs de transcription Drosophiles et mammifères. Nous avons comparé les performances d'un modèle PWM, d'un modèle de mélange de PWMs, et d'un modèle de Potts, en utilisant un critère bayésien (BIC) pénalisant les modèles à grand nombre de paramètres. Nous avons exhibé l'existence de corrélations faibles dont la prise en compte permet de significativement



**FIGURE 2.4 – Histogrammes des valeurs des champs  $h$  et couplages  $J$ .**

Histogrammes réalisés à partir des valeurs obtenues pour l'ensemble des TFs. Les champs et les couplages sont montrés en valeur absolue sur une échelle logarithmique d'espace-ment 0.05, et les valeurs nulles ne sont pas représentées. (A) Champs  $h_{inde}$  dans le modèle indépendant. (B) Champs  $h_{pw}$  et couplages  $J$  dans le modèle de Potts.

améliorer la description des données, le modèle de Potts étant significativement supérieur aux deux autres modèles dans la plupart des cas (22/28). Les interactions ont été étudiées systématiquement, montrant notamment une prépondérance des interactions entre plus proches voisins. Nous avons établi une correspondance entre les PWMs du modèle de mélange et les PWMs décrivant les états métastables du paysage énergétique généré par le modèle de Potts. Enfin, nous avons montré que les corrélations pouvaient être groupées en patterns de Hopfield

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

---

ou « mémoires », et qu'un petit nombre était suffisant à reconstruire le paysage d'interactions.

Une perspective intéressante de ce travail serait de conduire la même analyse sur des données grande échelle obtenues *in vitro* par la méthode HT-SELEX ([Jolma et al., 2013](#)). Notamment, certains des facteurs que nous avons étudiés *in vivo* sont représentés dans ces données, et il serait intéressant de voir les différences entre les modèles obtenus. Notamment, retrouve-t-on les mêmes corrélations ? Peut-on exhiber des spécificités de la fixation *in vivo*, où l'on s'attend à avoir des effets provenant de diverses sources (fixation de nucléosomes, superposition de sites de fixations, ...) ? Ces questions feront certainement l'objet d'un prochain travail.

---

## Chapitre 3

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

---

<b>3.1</b>	<b>Quelques approches existantes pour la recherche de motifs et de modules de régulation</b>	<b>84</b>
3.1.1	MEME : une approche <i>de novo</i> par Espérance-Maximisation . . . . .	85
3.1.2	STUBB : une méthode utilisant les corrélations entre sites de fixation et la phylogénie . . . . .	88
3.1.3	MONKEY : vers des modèles phylogénétiques plus complexes . . . . .	90
3.1.4	Approches sans motifs ou <i>motif-blind</i> . . . . .	93
3.1.5	Autres méthodes utilisant des collections d'oligonucléotides . . . . .	95
<b>3.2</b>	<b>Article</b>	<b>96</b>
<b>3.3</b>	<b>Calcul de la moyenne de la postérieure par une méthode MCMC</b>	<b>96</b>
3.3.1	Principe de l'algorithme de Metropolis-Hastings . . . . .	97
3.3.2	Application au calcul de la postérieure . . . . .	98
3.3.3	Illustration sur un exemple . . . . .	100
<b>3.4</b>	<b>Conclusion et perspectives</b>	<b>106</b>

## Introduction du chapitre 3

Dans le chapitre 2, nous avons vu comment décrire l'interaction TF-ADN lorsque des sites de fixation sont connus. Dans ce chapitre, nous adoptons une démarche plus générale. Nous connaissons l'activité de régulation d'un certain nombre de CRMs, et nous souhaitons savoir quels TFs s'y fixent (recherche de motifs), et si le génome contient d'autres CRMs avec la même activité (recherche de modules). Un algorithme permettant précisément de réaliser ces étapes a été développé précédemment par Hervé Rouault et appliqué au cas de la différenciation des organes sensoriels de la Drosophile (Rouault et al., 2010). Cet algorithme se distingue des précédents par le fait qu'il n'utilise pas de motifs connus en entrée mais les génère purement *de novo*, et par son utilisation systématique de l'information provenant de la conservation chez d'autres espèces grâce à des modèles d'évolution, le rendant notamment adapté au cas où les CRMs connus sont en petit nombre. Nous présentons ici Imogene, l'extension de cet algorithme au cas des mammifères, ainsi que son utilisation comme outil de classification de CRMs associés à différentes régulations.

Avant de rentrer dans le détail d'Imogene, nous présentons les méthodes existantes de recherche de motifs dans des CRMs. Le problème général est le suivant : étant données des CRMs conduisant à une même régulation (l'ensemble d'apprentissage), peut-on construire des modèles de sites de fixation qui « expliquent » cette co-régulation, c'est-à-dire qui prédisent l'existence de sites sur les CRMs mais pas sur des séquences ne participant pas à la co-régulation ?

### 3.1 Quelques approches existantes pour la recherche de motifs et de modules de régulation

Nous avons déjà introduit différentes méthodes de prédiction de motifs et modules en introduction (section 1.6). Ici nous décrivons plus en avant certaines de ces méthodes que nous jugeons utiles à la mise en perspective d'Imogene, soit par leur approche de génération *de novo* de motifs, soit par leur utilisation de la conservation chez d'autres espèces et de modèles d'évolution pour la prendre en compte, soit par le fait qu'elles développent des statistiques appropriées à l'étude de petits échantillons de CRMs. Pour une revue plus exhaustive, le lecteur intéressé pourra se référer à Wasserman and Sandelin (2004) et Aerts (2012).

### 3.1. Quelques approches existantes pour la recherche de motifs et de modules de régulation

#### 3.1.1 MEME : une approche *de novo* par Espérance-Maximisation

L'une des premières approches pour la prédiction *de novo* de motifs à partir de séquences a été celle de MEME ([Bailey and Elkan, 1994](#)), un algorithme basé sur la méthode d'Espérance-Maximisation ou EM (*Expectation Maximization*) utilisée précédemment dans ce cadre par [Lawrence and Reilly \(1990\)](#). Cet algorithme utilise une approche générative pour décrire les processus probabilistes qui ont permis la génération des séquences CRMs, ce qui permet d'écrire la probabilité qu'une séquence soit générée par un motif, et inversement de trouver le meilleur motif décrivant des séquences données. L'approche est illustrée en figure 3.1.

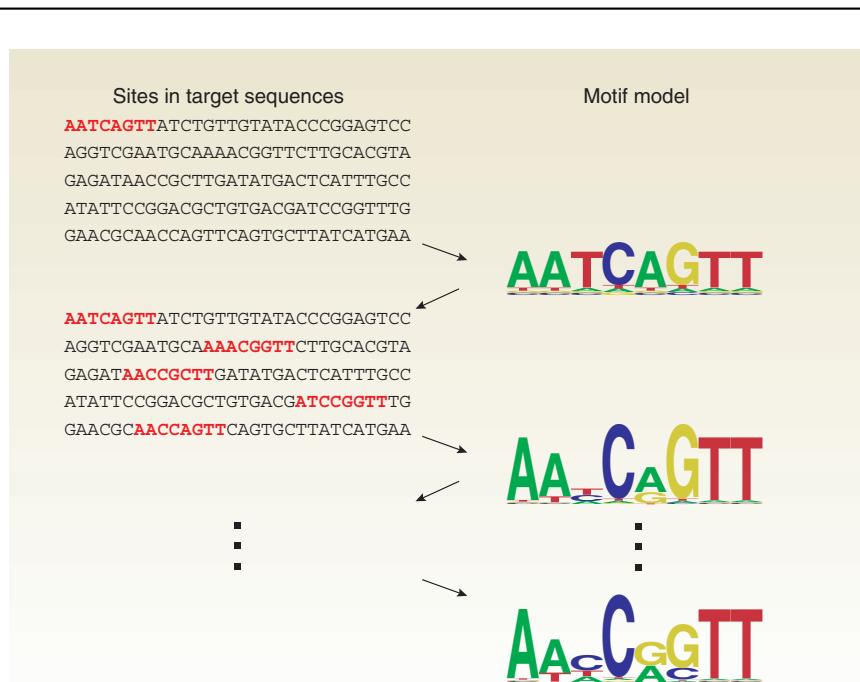


FIGURE 3.1 – Illustration de l'approche Espérance-Maximisation.

Figure tirée de [D'haeseleer \(2006\)](#) décrivant l'approche EM. Un premier modèle de motif est construit à partir d'un site initial. Ce modèle permet de pondérer l'ensemble des sites sur les séquences (étape E). En rouge sont montrés les meilleurs sites pour chaque séquence. En utilisant les poids des sites, il est possible de construire un modèle de vraisemblance maximale (étape M). La méthode originale de [Lawrence and Reilly \(1990\)](#) fait l'hypothèse qu'il y a exactement un site de fixation par séquence, condition qui est relâchée par MEME.

- **Vraisemblance d'une séquence**

Notons  $S = \{S_1, \dots, S_L\}$  une séquence<sup>15</sup> de taille  $L$ . Supposons qu'il y a exactement un site de régulation par CRM. C'est l'approche de [Lawrence and Reilly \(1990\)](#), et cette condition est relâchée par MEME, qui autorise l'utilisateur à préciser un nombre moyen de sites par séquence. La probabilité que la séquence possède un site de taille  $K$  à la position  $i$  étant donné le modèle de motif  $\mathcal{M}$  s'écrit

$$P(S|i, \mathcal{M}) = P_0(S_{1,i-1}) \times P(S_{i,i+K-1}|\mathcal{M}) \times P_0(S_{K,L}) \quad (3.1)$$

où  $S_{i,j}$  dénote la séquence entre les positions  $i$  et  $j$  incluses,  $P(S_{i,i+K-1}|\mathcal{M})$  est la probabilité de générer la séquence de taille  $K$  débutant à la position  $i$  avec le modèle  $\mathcal{M}$  (voir section [2.2](#)), et  $P_0(s)$  est la probabilité dite *background* de générer la séquence  $s$  étant donné un modèle génératif neutre, généralement pris comme étant une chaîne de Markov  $\mathcal{P}_k$  d'ordre  $k$  petit (0 à 2) :

$$P_0(S_{i,j}) = \prod_{l=i}^j \mathcal{P}_k(S_l|S_{l-k,l-1}) \quad (3.2)$$

L'équation [3.1](#) décrit donc la probabilité de générer la séquence  $S$  avec le modèle *background*, sauf à la position  $i$  où un site est généré avec le modèle de fixation  $\mathcal{M}$ . La probabilité de générer la séquence s'obtient finalement en sommant sur les positions pondérées par la probabilité *a priori*  $P(i)$  que le site soit à la position  $i$  :

$$P(S|\mathcal{M}) = \sum_{i=1}^{L-K+1} P(i)P(S|i, \mathcal{M}) \quad (3.3)$$

Cette probabilité est généralement prise uniforme, mais on peut y incorporer certaines informations, comme le nombre de séquences alignées (*reads*) d'une expérience de ChIP-seq.

- **Apprentissage du modèle**

Maintenant que nous savons exprimer la vraisemblance d'une séquence régulée par le motif  $\mathcal{M}$ , nous pouvons apprendre le meilleur modèle possible l'ayant générée : c'est la maximisation de la vraisemblance. Soit un ensemble de séquences  $S$  constitué de  $M$  séquences co-régulées  $S[1], \dots, S[M]$ . Ces séquences étant supposées indépendantes, la vraisemblance que ces données soient générées par un modèle  $\mathcal{M}$  est le produit sur les séquences de la

<sup>15</sup> On concatène les deux brins d'ADN dans cette séquence. On suppose en effet qu'ils participent équiprobalement à la fixation. La séquence génomique double brins est donc de longueur  $L/2$ .

---

3.1. Quelques approches existantes pour la recherche de motifs et de modules de régulation

---

quantité  $P(S[m]|\mathcal{M})$ . Il est plus utile dans ce cas de regarder la log-vraisemblance, s'écrivant alors comme une somme :

$$l(\mathcal{S}|\mathcal{M}) = \sum_{m=1}^M \log P(S[m]|\mathcal{M}) \quad (3.4)$$

Nous désirons obtenir le modèle  $\mathcal{M}$  maximisant cette quantité<sup>16</sup>. Nous ne connaissons pas les positions exactes des sites, qui sont des « variables cachées » et il n'existe pas de méthode d'estimation simple permettant de résoudre ce problème. C'est à ce stade qu'intervient la méthode Espérance-Maximisation (EM) (Dempster et al., 1977). L'algorithme EM est une méthode itérative qui part d'un modèle initial  $\mathcal{M}^0$  permettant de calculer les poids des positions dans les séquences (étape E d'espérance), puis estime le meilleur modèle  $\mathcal{M}^1$  étant données ces poids (étape M de maximisation). L'itération a lieu jusqu'à convergence vers un maximum local.

Notons  $\mathcal{M}^t$  le modèle à l'itération  $t$ . La probabilité qu'un site à la position  $i$  dans la séquence  $S[i]$  soit un site de fixation s'écrit  $P(i|S[m], \mathcal{M}^t)$ . On définit la log-vraisemblance moyenne d'un modèle  $\mathcal{M}$  à l'itération  $t$  par :

$$Q(\mathcal{M}|\mathcal{M}_t, \mathcal{S}) = \sum_m \sum_i P(i|S[m], \mathcal{M}^t) \log P(\mathcal{S}, i|\mathcal{M}) \quad (3.5)$$

Le modèle suivant  $\mathcal{M}^{t+1}$  est celui qui maximise cette quantité :

$$\mathcal{M}^{t+1} = \underset{\mathcal{M}}{\operatorname{argmax}} Q(\mathcal{M}|\mathcal{M}_t, \mathcal{S}) \quad (3.6)$$

L'équation 3.5 se scinde en une partie qui dépend de  $\mathcal{M}$  et une partie *background* qui n'en dépend pas (eq. 3.1), que l'on peut donc ignorer pour ce qui est de la maximisation. Ainsi, le modèle  $\mathcal{M}$  maximise la quantité suivante :

$$Q(\mathcal{M}|\mathcal{M}_t, \mathcal{S}) = \sum_m \sum_i P(i|S[m], \mathcal{M}^t) \log P(S_{i,i+K-1}|\mathcal{M}) \quad (3.7)$$

Chaque  $K$ -mer  $S_{i,i+K-1}$  des séquences de  $\mathcal{S}$  est donc pris en compte dans l'apprentissage en proportion de la croyance courante  $P(i|S[m], \mathcal{M}^t)$  que c'est un site de fixation.

Pour résumer, on a deux étapes :

---

16. La distribution *background* étant fixée (par exemple la chaîne de Markov peut être apprise sur un grand nombre de séquences intergéniques non codantes).

- étape E : utiliser  $\mathcal{M}^t$  pour attribuer un poids à chaque K-mer des séquences
- étape M : apprendre  $\mathcal{M}^{t+1}$  qui a la plus grande vraisemblance de générer les données pondérées par  $\mathcal{M}^t$ .

Reste le problème de choisir un modèle initial adéquat pour être sûrs de converger vers un maximum de vraisemblance global et pas juste local. MEME adopte pour cela une approche semi-exhaustive. Les différents K-mers des séquences d'apprentissage sont successivement utilisés pour générer un modèle initial. L'algorithme EM est itéré une fois. Le modèle de plus grande (log-)vraisemblance est finalement gardé comme motif initial pour une itération complète.

### 3.1.2 STUBB : une méthode utilisant les corrélations entre sites de fixation et la phylogénie

L'algorithme STUBB ([Sinha et al., 2003](#)) décrit les séquences par un modèle de Markov caché (HMM pour *Hidden Markov Model*) et les motifs par des PWMs. Il est basé sur l'algorithme Ahab ([Rajewsky et al., 2002](#)) – lui-même basé sur l'algorithme MobyDick ([Bussemaker et al., 2000](#)) –, qui peut être vu comme une extension de MEME au cas où les séquences contiennent plusieurs sites de fixations pour différents motifs. Notamment, ces méthodes ont l'intérêt tout comme MEME de ne pas avoir de seuil arbitraire pour définir un site car elles moyennent sur toutes les positions de sites (ou segmentations) possibles de la séquence. On parle de modèles thermodynamiques (voir [1.6.1](#)). La différence entre STUBB et Ahab est qu'il introduit deux informations supplémentaires : les corrélations entre motifs et la phylogénie. Enfin, contrairement à MEME, l'algorithme utilise comme condition initiale un ensemble de motifs connus pour être impliqués dans la co-régulation étudiée.

- **Description du modèle HMM**

Nous décrivons d'abord l'algorithme Ahab ([Rajewsky et al., 2002](#)). Notons  $W$  l'ensemble des motifs initiaux. Le modèle HMM à l'ordre 0 (HMM0) utilisé décrit la génération d'une séquence  $S$  de la manière suivante. La séquence est initialement de taille nulle. Le processus choisit un motif  $w_i \in W$  avec une probabilité  $p_i$  ou le motif *background*  $w_b$  (une PWM de longueur 1) avec une probabilité  $1 - \sum_i p_i$ . Une fois le motif  $w$  choisi, une séquence est échantillonnée à partir de la PWM de  $w$  et est ajoutée à la séquence  $S$ . Le processus est itéré jusqu'à ce que la séquence générée atteigne une taille  $L$ . La séquence de motifs choisis au cours de la

---

3.1. Quelques approches existantes pour la recherche de motifs et de modules de régulation

---

procédure définit une segmentation  $T$ . La probabilité que la séquence observée soit générée par ce processus de paramètres  $\theta = \{w_i, p_i\}$  est

$$P(S|\theta) = \sum_T P(T|\theta)P(S|T, \theta) \quad (3.8)$$

et peut être calculée par programmation dynamique (algorithme forward-backward). Le score d'une séquence est obtenu en comparant cette probabilité et la probabilité  $P(S|\theta_b)$  que la séquence soit générée uniquement par le modèle *background* :

$$F(S) = \operatorname{argmax}_{\theta} \log \left( \frac{P(S|\theta)}{P(S, \theta_b)} \right) \quad (3.9)$$

Le paramètre  $\theta$  (c'est-à-dire les  $p_i$ ) qui maximise le membre de droite est obtenu grâce à un algorithme de type EM ([Sinha et al., 2003](#)).

- **Ajout des corrélations entre motifs**

Des informations sur les corrélations entre motifs sont introduites dans  $\theta$  sous la forme de probabilités de transition  $p_{ij}$  que le motif choisi lors de la génération de la séquence soit  $w_j$  lorsque le premier motif précédent non-*background* est  $w_i$ . Parce que le nombre de paramètres devient grand, seules les corrélations importantes (dépassant un seuil fixé) sont ajoutées.

- **Incorporation de l'information phylogénétique**

Enfin, STUBB utilise l'information provenant de la conservation de la séquence chez d'autres espèces. Les séquences des différentes espèces sont d'abord alignées, puis la probabilité de générer l'alignement est calculée à l'aide d'un modèle phylogénétique. Ce modèle permet de prendre en compte le fait que les séquences homologues sont corrélées du fait qu'elles dérivent d'un ancêtre commun. Dans le cas de Stubb, les espèces sont supposées liées par un topologie en étoile, c'est-à-dire que les espèces partagent un seul ancêtre commun. Le modèle d'évolution suppose que les différentes bases de la séquence évoluent indépendamment, mutent à la même fréquence, et que la probabilité de fixation d'une mutation  $b \rightarrow b'$  à la position  $i$  est proportionnelle au poids  $w_{i,b'}$  de la PWM du nucléotide  $b'$  à cette position. Ce modèle est identique au modèle *Felsenstein* que nous introduisons dans l'article (voir section [3.2](#)) et qui est inspiré du modèle neutre de [Felsenstein \(1981\)](#) – les probabilités neutres étant remplacées par les fréquences PWM –. La probabilité  $P(\sigma|w)$  de générer l'alignement  $\sigma$  de séquences  $s$  avec le motif  $w$  de taille  $L_w$  s'écrit alors :

$$P(\sigma|w) = \prod_{i=1}^{L_w} \left[ \sum_b w_{i,b} \prod_{s \in \sigma} (q_s \delta_{b,s_i} + (1 - q_s) w_{i,s_i}) \right] \quad (3.10)$$

où  $w_{i,s_i}$  est la probabilité de générer le nucléotide  $s_i$  à la position  $i$  pour le motif  $w$ ,  $\delta_{x,y} = 1$  si  $x = y$  et 0 sinon, et  $q_s = e^{-\lambda t_s}$  est la probabilité de conserver un nucléotide au cours de l'évolution, qui est une fonction du taux de mutation neutre  $\lambda$  et du temps d'évolution  $t_s$  entre l'ancêtre commun et l'espèce  $s$ . En résumé, pour chaque position  $i$ , un nucléotide  $b$  est généré chez l'ancêtre commun avec une probabilité  $w_{i,b}$ , puis ce nucléotide est soit conservé chez l'espèce  $s$  avec une probabilité  $q_s$  ou bien il mute avec une probabilité  $1 - q_s$ , et une nouvelle base est sélectionnée selon les poids définis par  $w_i$ . Pour des espèces proches,  $q \sim 1$  et le fait d'observer des bases différentes à des positions homologues diminue fortement  $P(\sigma|w)$ , même si leur fréquence *a priori* donnée par  $w$  est identique : le modèle donne alors naturellement plus de poids aux séquences relativement conservées. Pour des espèces lointaines,  $q \sim 0$  et tout se passe comme si les séquences de  $\sigma$  étaient indépendantes.

### 3.1.3 MONKEY : vers des modèles phylogénétiques plus complexes

Dans la section précédente, le modèle phylogénétique utilisé par Stubb est relativement simple et la probabilité de fixation n'a pas de base claire. L'algorithme MONKEY ([Moses et al., 2004](#)) propose d'utiliser des modèles d'évolution plus complexes pour détecter les sites conservés à partir de motifs connus.

Les motifs  $w$  sont décrits par le modèle PWM et le *background* par les fréquences des nucléotides  $\pi_b$ . Le score d'un site est défini par la fraction des probabilités de générer la séquence  $s$  par l'un ou l'autre des modèles (*log-likelihood ratio* ou LLR) :

$$LLR(s) = \log \frac{P(s|w)}{P(s|\pi)} = \sum_i \log \frac{w_{i,b(i)}}{\pi_b} \quad (3.11)$$

Le but est de généraliser ce score au cas d'un alignement  $\sigma$  de séquences  $s$ , en réalisant son calcul sur l'ancêtre commun des séquences. Cet ancêtre commun, ainsi que tous les ancêtres communs intermédiaires pour une topologie d'arbre  $T$  quelconque, ne sont pas observés, et il faut donc sommer sur tous leurs états (nucléotides) possibles étant donnés l'alignement observé  $\sigma$ , l'arbre  $T$  et le modèle d'évolution. La solution générale de ce problème a été donnée par [Felsenstein \(1981\)](#). Notamment, nous pouvons nous concentrer sur le cas à deux espèces,

### 3.1. Quelques approches existantes pour la recherche de motifs et de modules de régulation

---

le cas général procédant par récurrence à partir de ce cas simple. Nous pouvons aussi nous concentrer sur une position  $i$  donnée, puisque les positions sont indépendantes.

Considérons un alignement de deux nucléotides  $s_1$  et  $s_2$ . On définit un nouveau score comme étant le LLR comparant l'hypothèse que  $s_1$  et  $s_2$  représentent un site conservé pour un motif  $w$  et l'hypothèse que les bases ont été tirées dans le *background* :

$$LLR_{\text{cons}}(s_1, s_2) = \log \frac{P(s_1, s_2 | w, T, R_w)}{P(s_1, s_2 | \pi, T, R_{\text{back}})} \quad (3.12)$$

où  $R_w$  et  $R_{\text{back}}$  sont des matrices de taux de transition décrivant les processus de substitution au cours de l'évolution pour le cas d'un site de fixation du motif  $w$  et pour le *background*, et interviennent dans l'écriture de la probabilité de transition de la base  $b$  à la base  $b'$  :

$$p_{b \rightarrow b'} = \left( e^{R_M d} \right)_{b, b'} \quad (3.13)$$

où  $R_M$  est la matrice de taux de transition du modèle  $M$  et  $d$  le temps évolutif. Dans le cas simple à deux espèces, l'arbre  $T$  est en étoile, avec des distances  $d_1$  et  $d_2$  de l'ancêtre commun aux espèces contenant les bases  $s_1$  et  $s_2$ . Par ailleurs, les espèces évoluent indépendamment depuis leur séparation. Notant  $b$  la base sur l'ancêtre commun, on a finalement :

$$\begin{aligned} P(s_1, s_2 | M, T, R_M) &= \sum_b P(s_1 | b, d_1, R_M) P(s_2 | b, d_2, R_M) P(b | M) \\ &= \sum_b \left( e^{R_M d_1} \right)_{b, s_1} \left( e^{R_M d_2} \right)_{b, s_2} P(b | M) \end{aligned} \quad (3.14)$$

Le calcul sur un nombre quelconque d'espèces se fait récursivement ([Felsenstein, 1981](#)), jusqu'à la racine de l'arbre où  $P(b | M)$  vaut  $w_{i,b}$  pour le motif et  $\pi_b$  pour le *background*.

Plusieurs modèles peuvent être choisis pour les matrices de transition. Dans le cas du *background*, il peut être décrit par des modèles neutres. Par exemple le modèle de Felsenstein ([Felsenstein, 1981](#)) est le modèle le plus simple dont la distribution d'équilibre redonne les fréquences du *background*. Le modèle HKY ([Hasegawa et al., 1985](#)) est une variante qui inclut le fait que les mutations entre bases de même nature chimique (purine ou pyrimidine), appelées transitions, sont 2 fois plus fréquentes que les autres mutations, appelées transversions. Ce dernier modèle est utilisé dans MONKEY pour décrire le *background*. Pour ce qui est du motif, les taux de transition dépendent de la position au sein du site : par exemple les bases

dégénérées mutent plus vite que les bases très conservées (Moses et al., 2003). Pour prendre cette variation en compte, une possibilité est de modifier le modèle neutre en utilisant à la place des fréquences *background* les fréquences données par la PWM : c'est ce qui est fait dans Stubb avec le modèle Felsenstein (éq.3.10). Dans MONKEY, les auteurs utilisent un modèle plus complexe, appelé modèle Halpern-Bruno ou HB, préalablement introduit pour étudier l'évolution des régions codantes (Halpern and Bruno, 1998). Dans ce modèle, le taux de substitution  $R(i)_{a,b}$  de la base  $a$  vers la base  $b$  en position  $i$  est à une constante près le produit de la probabilité de mutation neutre de  $a$  vers  $b$  (indépendante de la position) par une probabilité de fixation  $f_{a,b}^i$  (dépendante de la position) :

$$R(i)_{a,b} \propto Q_{a,b} f_{a,b}^i \quad (3.15)$$

où  $Q = R_{\text{back}}$  est la matrice de transition du modèle *background*. La probabilité de fixation peut être obtenue en utilisant un résultat classique de génétique des populations (Kimura, 1962) :

$$f_{a,b}^i \simeq \frac{2s}{1 - e^{-2Ns}} \quad (3.16)$$

$$f_{b,a}^i \simeq \frac{-2s}{1 - e^{2Ns}} \quad (3.17)$$

où  $N$  est la taille effective de la population (le facteur 2 vient du fait que la population est diploïde),  $s$  est la valeur adaptative relative ou *fitness* de la base  $b$  par rapport à la base  $a$  en position  $i$ , et l'évolution est quasi-neutre ( $s \ll 1$ ). Cette dernière hypothèse permet d'écrire :

$$\frac{f_{a,b}^i}{f_{b,a}^i} \simeq e^{2Ns} \quad (3.18)$$

Par ailleurs, en supposant que les substitutions vérifient le bilan détaillé à l'équilibre, on peut écrire

$$\frac{f_{a,b}^i}{f_{b,a}^i} = \frac{w_{i,b} Q_{b,a}}{w_{i,a} Q_{a,b}} \quad (3.19)$$

Finalement, en combinant les équations 3.16, 3.18 et 3.19, on obtient la matrice de transition du modèle HB en position  $i$  sous la forme :

$R(i)_{a,b} \propto Q_{a,b} \frac{x \log x}{x - 1}$

(3.20)

---

 3.1. *Quelques approches existantes pour la recherche de motifs et de modules de régulation*


---

où

$$x = \frac{w_{i,b}Q_{b,a}}{w_{i,a}Q_{a,b}} \simeq e^{2Ns} \quad (3.21)$$

traduit l'effet de la sélection. En développant autour de  $x = 1$  et en restant au premier ordre, on a

$$\frac{x \log x}{x - 1} \simeq \frac{1}{2}(1 + x) \quad (3.22)$$

Ainsi, dans le cas neutre où  $x = 1$ , la matrice de transition se réduit à la matrice *background* :  $R(i)_{a,b} = Q_{a,b}$ . Cependant, lorsque la base  $b$  est plus conservée que la base  $a$  ( $x > 1$ ), les substitutions de  $a$  vers  $b$  sont plus fréquentes que sous le modèle neutre :  $R(i)_{a,b} > Q_{a,b}$ .

### 3.1.4 Approches sans motifs ou *motif-blind*

Les algorithmes précédents utilisent en leur cœur un modèle de motif  $\mathcal{M}$ , généralement une PWM, permettant d'attribuer une probabilité  $P(S|\mathcal{M})$  à une séquence donnée. Néanmoins, il existe certaines méthodes cherchant à décrire plus généralement la statistique des mots au sein des CRMs sans chercher à associer ces statistiques à des motifs ayant une caractérisation biochimique précise. De telles approches sont dites sans motifs (*motif-blind*). Nous en recensons ici quelques-unes (voir [Kantorovitz et al. \(2009\)](#) pour plus de détails).

- **Modèles basés sur des chaînes de Markov**

Plusieurs modèles basés sur des chaînes de Markov ont été proposés. Par exemple, l'algorithme PFRSampler de [Grad et al. \(2004\)](#) consiste en un apprentissage de modèles de Markov d'ordre 5 sur des séquences d'intérêt et sur des séquences *background*, ces séquences étant préalablement filtrées par la conservation phylogénétique. Il est ensuite possible de calculer la vraisemblance qu'une séquence donnée soit générée par l'un ou l'autre des modèles, de manière similaire à l'éq.[3.3](#). Le score d'une séquence  $S$  de taille  $L$  est défini comme étant la différence des log-vraisemblances qu'elle soit générée par le modèle d'intérêt  $\mathcal{M}_{\text{train}}$  et par le modèle *background*  $\mathcal{M}_{\text{back}}$  :

$$\text{Score}(S) = \log \frac{P(S|\mathcal{M}_{\text{train}})}{P(S|\mathcal{M}_{\text{back}})} = \sum_{i=1}^L \log \frac{T_{\text{train}}(S_i|S_{i-k,i-1})}{T_{\text{back}}(S_i|S_{i-k,i-1})} \quad (3.23)$$

où  $T_{\text{train}}$  et  $T_{\text{back}}$  sont les probabilités de transition associées aux deux modèles,  $S_{i,j}$  est la séquence entre les positions  $i$  et  $j$  incluses, et  $k$  est l'ordre de la chaîne de Markov (ici  $k = 5$ ).

Cette méthode détecte donc la *signature* globale d'un CRM plutôt que la présence de sites de fixation pour des TFs particuliers. Cette méthode a aussi été implémentée par [Ivan et al. \(2008\)](#) sous le nom de *Markov Chain Discrimination* (MCD), avec la différence notable que les auteurs n'utilisent pas la phylogénie. Une généralisation de cette approche a été proposée par [Kazemian et al. \(2011\)](#) sous le nom d'*Interpolated Markov Model*. Au lieu d'utiliser une chaîne de Markov à un ordre donné, les auteurs réalisent une interpolation entre des chaînes de Markov d'ordres 0 à 5, en ne gardant pour chaque ordre que les transitions sur-représentées dans les séquences d'apprentissage. Ceci leur permet de capturer les signatures présentes à différentes résolutions.

- **Modèles basés sur des enrichissements en  $k$ -mers**

D'autres modèles sont basés sur la statistique des mots de  $k$  nucléotides ( $k$ -mers) dans les séquences d'apprentissage. Par exemple, [Kantorovitz et al. \(2007\)](#) ont introduit une mesure de similarité entre séquences basée sur le nombre de  $k$ -mers qu'elles ont en commun. Les auteurs définissent le score  $D_2$  par

$$D_2(S_1, S_2) = \sum_{\{w\}} N_1(w)N_2(w) \quad (3.24)$$

où  $S_i$  est la séquence  $i$ ,  $\{w\}$  est l'ensemble des  $k$ -mers, et  $N_i(w)$  est le nombre de  $k$ -mers  $w$  dans la séquence  $i$ . Ce score est grand si les séquences partagent de nombreux  $k$ -mers, c'est-à-dire si elles ont une régulation commune. Ce score est normalisé pour produire le  $z$ -score (c'est-à-dire le nombre d'écart-type par rapport à la moyenne) de mesure de similarité  $D2z$  :

$$D2z(S_1, S_2) = \frac{D_2(S_1, S_2) - E(D_2)}{\sigma(D_2)} \quad (3.25)$$

où  $E(D_2)$  et  $\sigma(D_2)$  sont l'espérance et l'écart-type de la distribution de  $D_2(S_1, S_2)$ , calculés théoriquement sous l'hypothèse que les séquences  $S_1$  et  $S_2$  sont indépendantes et sont générées par un modèle *background* de type chaîne de Markov.

D'autres méthodes pour attribuer un score à une séquence par similarité de  $k$ -mers avec des séquences d'apprentissage ont été introduites par [Kantorovitz et al. \(2009\)](#). Étant données des séquences d'apprentissage, les 200  $k$ -mers ( $k = 6$ ) les plus représentés par rapport à un modèle *background* sont sélectionnés selon leur  $z$ -score, dans ce cas le nombre d'écart-type séparant le nombre  $n(w)$  de fois que le mots apparaît dans le training set du nombre de fois moyen  $\lambda(W)$  qu'il devrait apparaître sous un modèle *background* ([Sinha and Tompa, 2000](#)).

---

### 3.1. Quelques approches existantes pour la recherche de motifs et de modules de régulation

---

Étant donnés ces mots sur-représentés, il est possible de définir un score basé sur la statistique de Poisson (modèle PAC pour *Poisson Additive Conditional*) :

$$PAC(S) = \frac{1}{200} \sum_w F(\lambda(w), n(w) - 1) \quad (3.26)$$

où  $F(\lambda, x)$  est la distribution de Poisson cumulative de paramètre  $\lambda$ , donnant une valeur faible (proche de 0) si  $n(w) \simeq \lambda(w)$  et maximale (proche de 1) si  $n(w) \gg \lambda(w)$ . D'autres scores sont définis par une approche de classification linéaire pondérant les comptages de  $k$ -mers (WSC pour *Weighted Sum of Counts*)

$$WSC(S) = \sum_w \beta(w)n(w) \quad (3.27)$$

où  $\beta(w)$  est un poids reflétant l'association avec l'ensemble d'apprentissage. Ce poids peut être le rapport de la fréquence du mot dans l'ensemble d'apprentissage et de sa fréquence dans le *background* (modèle HexDiff, [Chan and Kibler \(2005\)](#)), le logarithme de cette quantité ([Rouault et al., 2010](#)), ou encore le  $z$  score introduit précédemment mesurant la sur-représentation du  $k$ -mer dans l'ensemble d'apprentissage (méthode HexYMF, [Kantorovitz et al. \(2009\)](#)).

### 3.1.5 Autres méthodes utilisant des collections d'oligonucléotides

Alors que les méthodes basées sur l'enrichissement en  $k$ -mers présentées en [3.1.4](#) s'intéressent au contenu général d'un CRM en  $k$ -mers, d'autres méthodes tentent de regrouper les  $k$ -mers en groupes associés à un régulateur putatif. Par exemple, [Cao et al. \(2010a\)](#) ont introduit un algorithme de recherche de motifs destiné à l'étude de données ChIP-seq. Ici, un motif est simplement défini comme une collection de  $k$ -mers. Le but est de trouver les motifs qui discriminent le mieux un ensemble de séquences positives (des pics de ChIP-seq) d'un ensemble de séquences *background*. L'algorithme énumère d'abord tous les  $k$ -mers, mesure leurs fréquences, et ajuste pour chacun un modèle de régression logistique mesurant sa capacité à classifier les séquences. Le  $k$ -mer le plus important est choisi comme graine. Puis toutes les variations à distance de Hamming de 1 et 2 (c'est-à-dire ayant un ou deux nucléotides différents) de cette graine sont énumérées, et sont ajoutées au motif si elles permettent d'améliorer la régression. Lorsqu'un motif final est obtenu, toutes ses occurrences sont masquées et un nouveau motif est appris. Un algorithme similaire, HOMER, a été développé par [Heinz et al.](#)

(2010). La différence majeure est que HOMER utilise la collection de  $k$ -mers obtenue pour générer une PWM qui est ensuite raffinée sur les séquences.

## 3.2 Article

Dans l'article suivant, nous introduisons Imogene, un algorithme de génération de motifs *de novo* utilisant la phylogénie basé sur l'algorithme de Rouault et al. (2010) qu'il généralise au cas des mammifères. Plusieurs tests sont réalisés, montrant sa capacité à prédire des CRMs tissu-spécifiques ou encore à classer différents CRMs selon leur motif d'expression.

## 3.3 Calcul de la moyenne de la postérieure par une méthode MCMC

Nous avons voulu savoir si l'approximation réalisée par Imogene lors du calcul du maximum de l'estimateur de la postérieure modifiée (obtenue avec un prior de Dirichlet de paramètres  $\alpha_b + 1$ ) donnait un résultat effectivement proche du calcul de la moyenne de l'estimateur de la postérieure non modifiée (avec un prior de Dirichlet de paramètres  $\alpha_b$ ) :

$$\operatorname{argmax}_{w_i} \mathcal{L}(\{\mathcal{A}\}|w_i) \operatorname{Dir}(w_i, \alpha_b + 1) \simeq \int w_i \mathcal{L}(\{\mathcal{A}\}|w_i) \operatorname{Dir}(w_i, \alpha_b) dw_i \quad (3.28)$$

où  $w_i$  est le vecteur de poids de la PWM à la position  $i$ ,  $\{\mathcal{A}\}$  est l'ensemble des alignements de nucléotides observés à cette position dans les sites de fixation,  $\mathcal{L}(\{\mathcal{A}\}|w_i)$  dénote la vraisemblance de l'alignement connaissant le vecteur de poids  $w_i$  et  $\operatorname{Dir}(w_i, \alpha_b)$  est le prior de Dirichlet de paramètres  $\alpha_b$ . Pour des distances phylogénétiques nulles ou infinies, cette approximation est valide, et pour des distances intermédiaires, elle semble fonctionner dans le cas simplifié de l'arbre en étoile de la figure S5 de l'article. Qu'en est-t-il dans le cas réel avec un arbre à la topologie plus complexe ? L'estimation directe de la moyenne de cette distribution est difficile, puisque nous n'avons pas de moyen simple de l'échantillonner. Afin de contourner ce problème, nous avons eu recours à une méthode de Monte-Carlo par chaînes de Markov ou MCMC basée sur l'algorithme de Metropolis-Hastings (Krauth, 2006). Nous présentons la comparaison de la moyenne de la postérieure obtenue par l'approche MCMC et du maximum de la postérieure modifiée par descente de gradient sur un cas réel simple en utilisant l'arbre des vertébrés (figure 2 de l'article) et le modèle d'évolution *Felsenstein*.

### 3.3.1 Principe de l'algorithme de Metropolis-Hastings

L'algorithme de Metropolis-Hastings ([Metropolis et al., 1953; Hastings, 1970](#)) permet d'échantillonner une distribution donnée en utilisant le parcours d'une chaîne de Markov ayant cette distribution pour loi stationnaire. Un tel processus de Markov est défini par des probabilités de transition  $P(w \rightarrow w')$  entre deux états  $w$  et  $w'$ . Il converge vers une distribution stationnaire  $\pi(w)$  unique sous deux conditions : (1) les transitions sont réversibles et le processus satisfait le bilan détaillé  $\pi(w)P(w \rightarrow w') = \pi(w')P(w' \rightarrow w)$ , (2) le processus est ergodique, c'est-à-dire que tout état est et reste accessible. L'algorithme de Metropolis-Hastings repose sur la construction d'une chaîne de Markov ayant ces propriétés et dont la distribution d'équilibre  $\pi(w)$  est la probabilité que l'on cherche à échantillonner  $P(w)$ . Pour cela, on part de l'équation du bilan détaillé, que l'on peut écrire

$$\frac{P(w \rightarrow w')}{P(w' \rightarrow w)} = \frac{P(w')}{P(w)} \quad (3.29)$$

La transition  $P(w \rightarrow w')$  est ensuite décomposée en deux sous-étapes, la proposition (*proposal*) et l'acceptation (*acceptance*) :

$$P(w \rightarrow w') = \underbrace{g(w \rightarrow w')}_{\text{proposition}} \cdot \underbrace{A(w \rightarrow w')}_{\text{acceptation}} \quad (3.30)$$

En insérant dans l'éq. 3.29 on obtient

$$\frac{A(w \rightarrow w')}{A(w' \rightarrow w)} = \frac{P(w')}{g(w \rightarrow w')} \frac{g(w' \rightarrow w)}{P(w)} \quad (3.31)$$

Plusieurs choix de la fonction d'acceptation sont possibles pour satisfaire cette équation ([Hastings, 1970](#)). Un choix courant, dit choix de Metropolis, est :

$$A(w \rightarrow w') = \min \left( 1, \frac{P(w')}{g(w \rightarrow w')} \frac{g(w' \rightarrow w)}{P(w)} \right) \quad (3.32)$$

On remarque que cette quantité est invariante sous multiplication de la distribution  $P(w)$  par un facteur non nul. Autrement dit, la distribution n'a pas besoin d'être normalisée. Dans un cadre bayésien, cela veut dire que l'on peut remplacer la postérieure par le produit de la vraisemblance et du *prior*.

La méthode de Metropolis-Hastings se résume donc ainsi :

1. Initialiser  $w$  à une certaine valeur.
2. Choisir un nouvel état  $w'$  tiré selon  $g(w \rightarrow w')$
3. Accepter l'état avec une probabilité donnée par  $A(w \rightarrow w')$ . Si le nouvel état n'est pas accepté, alors  $w' = w$ .
4. Itérer jusqu'à convergence

Au final,  $w$  étant tiré selon la distribution  $P(w)$ , sa moyenne est estimée en sommant les poids  $w(t)$  obtenus au cours des  $N$  itérations réalisées :

$$\langle w \rangle \simeq \hat{w}_N = \frac{1}{N} \sum_{t=1}^N w(t) \quad (3.33)$$

Quant au critère de convergence, une possibilité est d'utiliser le Théorème Central Limite (TCL). Celui-ci stipule que la moyenne de  $n$  variables aléatoires indépendantes et identiquement distribuées selon une loi de moyenne  $\mu$  et d'écart-type  $\sigma$  tous deux de valeurs finies suit, pour  $n$  grand, une loi normale de moyenne  $\mu$  et d'écart-type  $\sigma / \sqrt{n}$ . Dans l'approche MCMC, les échantillons successifs  $w_i$  ne sont pas indépendants à cause du fait qu'on les tire selon la loi  $g(w \rightarrow w')$ . Il faut donc calculer le temps de décorrélation  $T$  pour lequel  $\langle w(t)w_i(t+T) \rangle \simeq 0$ , puis utiliser les échantillons  $w(t)$  obtenus toutes les  $T$  itérations comme variables indépendantes. L'application du TCL permet alors d'arrêter les itérations lorsqu'une certaine précision désirée est atteinte, par exemple lorsque  $\sigma / \sqrt{n} < 0.05 \cdot \mu$ , c'est-à-dire lorsque les variations de l'estimateur de la moyenne sont de l'ordre de 5% de la valeur de la moyenne. L'écart-type  $\sigma$  étant lui-même estimé à partir des échantillons de l'algorithme, il faut aussi s'assurer qu'il a convergé vers une valeur stable pour appliquer le TCL.

### 3.3.2 Application au calcul de la postérieure

Dans notre cas, nous souhaitons utiliser l'algorithme de Metropolis-Hastings pour calculer la valeur moyenne du vecteur de poids  $w_i$  en position  $i$  de la PWM selon la distribution postérieure  $\mathcal{P}(w_i | \{\mathcal{A}\})$  :

$$\langle w_i \rangle = \int w_i \mathcal{P}(w_i | \{\mathcal{A}\}) dw \quad (3.34)$$

Il nous faut pour cela définir une loi de proposition  $g(w_i \rightarrow w'_i)$  pertinente. Dans notre cas, les poids  $w_i$  doivent rester dans le simplexe de dimension 3 défini par  $w_A, w_C, w_G > 0$  et

---

3.3. Calcul de la moyenne de la postérieure par une méthode MCMC

---

$w_A + w_C + w_G < 1$ , le poids  $w_T$  étant entièrement déterminé par  $w_T = 1 - w_A - w_C - w_G$ . La distribution naturelle possédant cette propriété est la loi de Dirichlet  $\text{Dir}(\alpha)$ , de paramètres  $\alpha = \{\alpha_A, \alpha_C, \alpha_G, \alpha_T\}$  et de densité de probabilité

$$f(w) = \frac{1}{B(\alpha)} \prod_{b \in \{A, C, G, T\}} w_{i,b}^{\alpha_b - 1} \quad (3.35)$$

où  $B(\alpha)$  est la fonction bêta multinomiale permettant la normalisation. Cette distribution est la même que celle obtenue dans le cas d'observations indépendantes (cf article). Afin d'accélérer l'échantillonnage MCMC, nous avons cherché à régler les paramètres  $\alpha$  de manière à être au plus proche de la distribution  $\mathcal{P}(w_i | \{\mathcal{A}\})$ . Dans le cas de  $N$  sites indépendants, celle-ci suit une loi de Dirichlet de paramètres  $\alpha_p + N_i$ , où  $\alpha_p$  est le vecteur de pseudo-counts et  $N_i$  le vecteur donnant les nombres d'observations des nucléotides en position  $i$  au sein des différentes séquences. Dans le cas d'un arbre phylogénétique corrélant les séquences, le nombre *effectif* d'observation est moins grand que le nombre total de séquences. Nous avons donc défini les paramètres de notre proposition comme étant

$$\alpha_b = \alpha_p + N_{\text{eff}} \cdot w_i \quad (3.36)$$

où  $N_{\text{eff}} = N_{\text{sites}} \cdot N_{\text{spe}} / 2$  avec  $N_{\text{sites}}$  le nombre d'alignements observés et  $N_{\text{spe}}$  le nombre d'espèces dans l'alignement (12 dans les deux cas, Drosophile et mammifères). Grossièrement, cela revient à dire que le modèle d'évolution réduit d'un facteur 2 le nombre de séquences indépendantes. Nous avons calculé que le taux d'acceptation (proportion de mouvements proposés qui sont acceptés) pour ce paramètre était de l'ordre de 50%, une valeur généralement considérée comme raisonnable (Krauth, 2006). Nous obtenons finalement l'expression pour la proposition :

$$g(w \rightarrow w') = \frac{1}{B(\alpha)} \prod_{b \in \{A, C, G, T\}} (w'_{i,b})^{\alpha_{p,b} + N_{\text{eff}} \cdot w_{i,b} - 1} \quad (3.37)$$

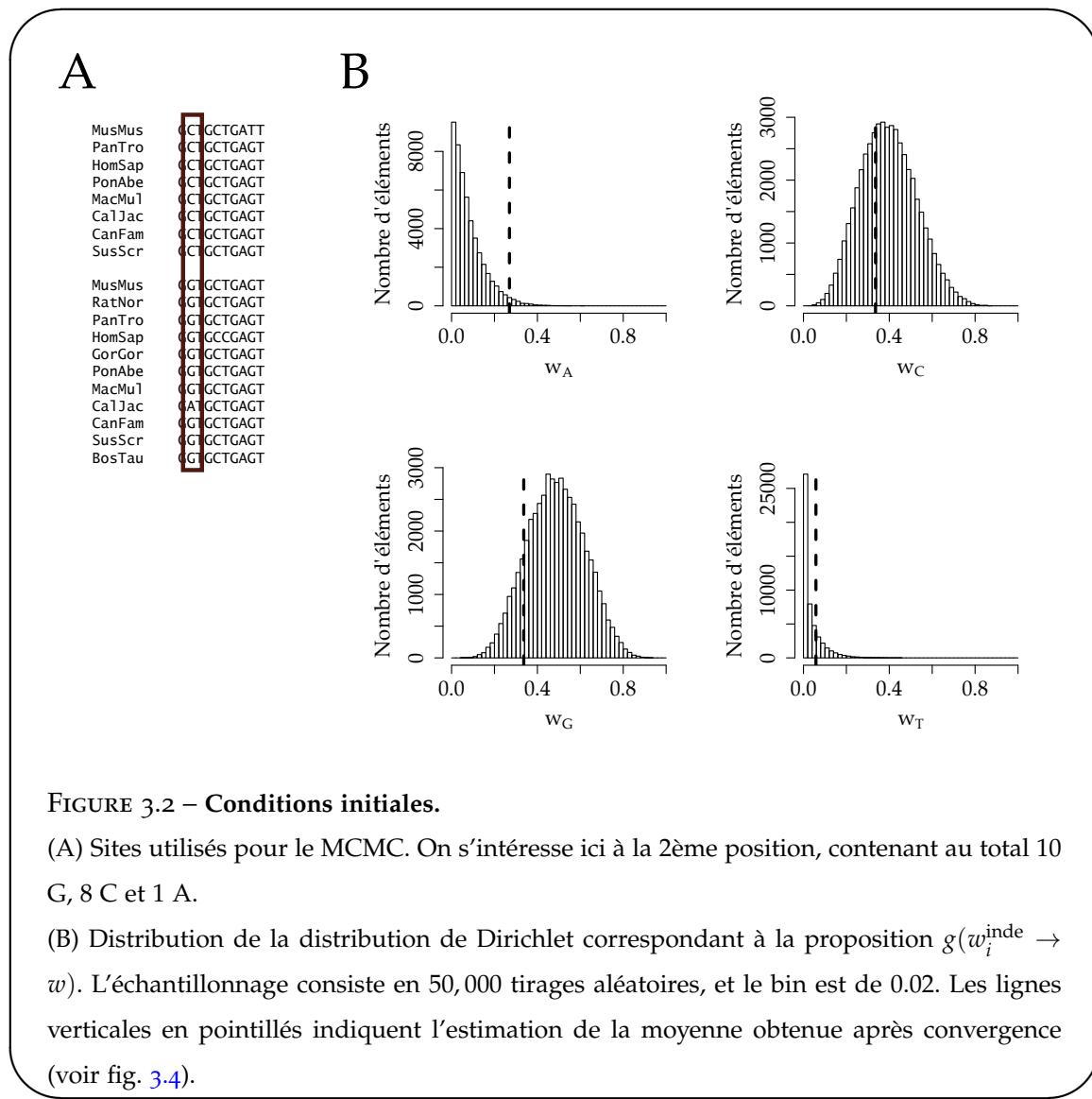
Le vecteur  $w_i$  est initialisé à la valeur qu'il prendrait si toutes les séquences orthologues étaient des observations indépendantes (cf article) :

$$w_{i,b}(0) = w_{i,b}^{\text{inde}} = \frac{N_{i,b} + \alpha_b}{N_{\text{tot}} + \sum_b \alpha_b} \quad (3.38)$$

Le poids  $w_i(1)$  suivant est tiré selon la probabilité de transition  $g(w_i(0) \rightarrow w_i(1))$ . Les différentes quantités de l'équation 3.32 sont ensuite calculées et la transition est acceptée avec

probabilité  $A$  ( $w_i(0) \rightarrow w_i(1)$ ).

### 3.3.3 Illustration sur un exemple



---

3.3. *Calcul de la moyenne de la postérieure par une méthode MCMC*


---

de la postérieure obtenue après convergence de la chaîne (voir ci-dessous) est aussi montrée : elle est relativement proche de la moyenne de la distribution, indiquant que le choix de la valeur initiale est effectivement judicieux.

La chaîne MCMC est ensuite lancée. Le taux d'acceptation est calculé comme valant 62%. Les 500 premiers échantillons de  $w_i$  sont montrés en figure 3.3A. On note que certains points sont corrélés : diminutions ou augmentations successives de la valeur courante  $w_i(t)$  sur plusieurs itérations. Pour quantifier cet effet, nous avons mesuré la corrélation temporelle des échantillons. Celle-ci est donnée par

$$C_b(\tau) = \frac{1}{N} \sum_{t=1}^{N-\tau} w_{i,b}(t)w_{i,b}(t+\tau) - \left( \frac{1}{N} \sum_{t=1}^N w_{i,b}(t) \right)^2 \quad (3.39)$$

avec dans ce cas  $N = 50,000$ . Le logarithme de cette quantité est montré en figure 3.3B. L'intérêt du logarithme est de mettre en exergue le caractère exponentiel de la décorrélation :

$$C_b(\tau) \propto e^{-\tau/T_b} \quad (3.40)$$

Les temps de décorrélation  $T_b \sim 10$  sont estimés en ajustant une droite sur les 20 premiers points de la courbe. Maintenant que l'on connaît le temps de corrélation entre deux échantillons, il est possible d'obtenir des échantillons indépendants en les choisissant à des intervalles plus grands que  $T_b$ . Dans notre cas, nous avons choisis un intervalle de 30 itérations.

Nous souhaitons maintenant étudier la convergence de la chaîne MCMC. Pour cela, nous utilisons le Théorème Central Limite (TCL, cf 3.3.1). Pour un nombre  $n$  suffisamment grand d'échantillons indépendants, l'écart-type de la distribution de l'estimateur empirique de la moyenne  $\hat{w}_n$  se comporte comme  $\sigma_b / \sqrt{n}$ , où  $\sigma_b$  est l'écart-type de la distribution  $\mathcal{P}(w_i | \mathcal{A})$ . Ce dernier doit lui-même être estimé à partir de la chaîne MCMC. Nous présentons en figure 3.4A la valeur de l'estimation  $\sigma_b(n)$  obtenue pour  $n$  itérations indépendantes. On voit que cette valeur atteint rapidement en  $\sim 100$  itérations une valeur stable, et que dans tous les cas les fluctuations sont faibles, de l'ordre de 10% de  $\hat{w}_n$ . Il paraît donc raisonnable d'utiliser cette valeur de  $\sigma_b$  pour le TCL. Nous traçons en figure 3.4B la quantité  $\sigma_b(n) / \sqrt{n}$ . La chaîne est considérée comme convergée lorsque cette valeur est inférieure ou égale à 5% de la valeur moyenne estimée  $\hat{w}_n$  (ligne grise) dans les 4 cas. Ce seuil est bien entendu arbitraire et dépend

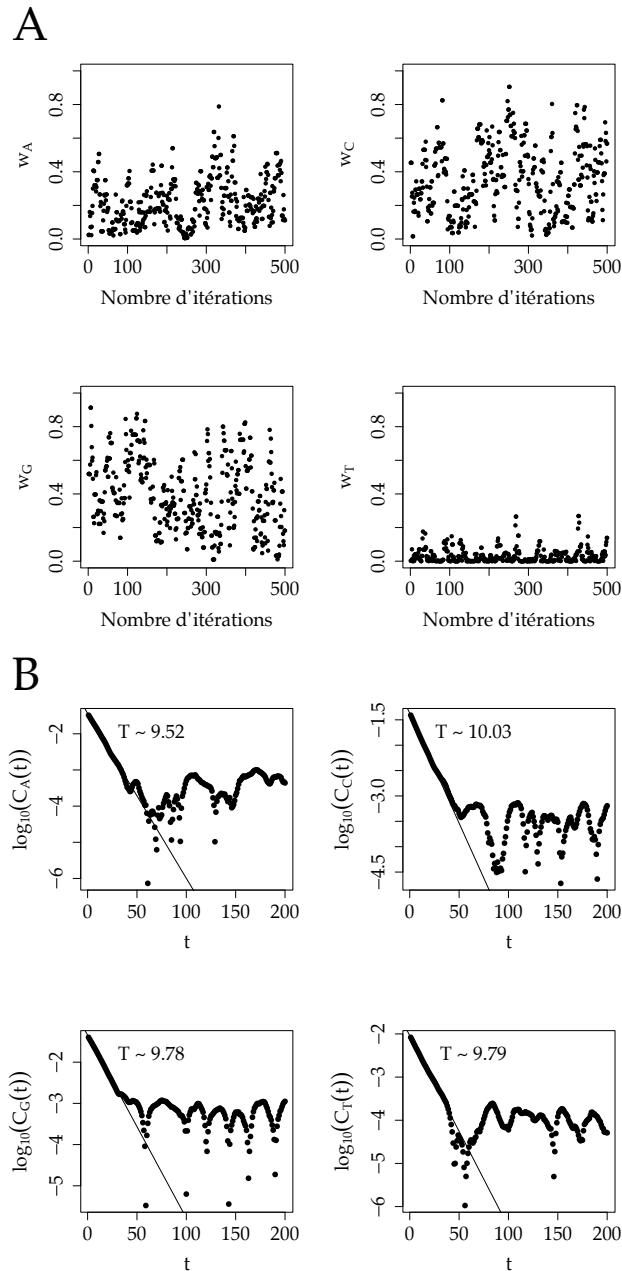


FIGURE 3.3 – Corrélations entre les échantillons.

- (A) Extrait de l'échantillonnage de  $w_i$ . Les 500 premiers échantillons  $w_i$  de la chaîne MCMC sont montrés.
- (B) Fonction d'auto-corrélation de la chaîne MCMC. La corrélation  $C_b(t)$  est réalisée entre des échantillons séparés de  $t - 1$  itérations. Une droite est ajustée sur les 20 premiers points, permettant d'obtenir la pente  $-T_b$  donnant le temps de décorrélation.

---

*3.3. Calcul de la moyenne de la postérieure par une méthode MCMC*

de la précision voulue par l'utilisateur sur l'estimation. Néanmoins, une plus grande précision implique un plus long temps de calcul.

Nous comparons en figure 3.5 la valeur de l'estimation du Maximum A Posteriori (MAP) de la postérieure modifiée (voir article) obtenue avec la méthode de descente de gradient et celle de la moyenne de la postérieure obtenue avec l'approche MCMC, en fonction du nombre total d'itérations. L'approche MCMC converge vers un état proche de celui donné par la descente de gradient. On note que la convergence de l'approche MCMC est beaucoup plus lente : plus de 10,000 itérations, alors que la descente de gradient n'en requiert que 100. Au vu de la faible différence entre les deux résultats, nous utilisons dans Imogene l'approche de maximisation de la postérieure modifiée, ce qui permet un gain de temps considérable pour l'algorithme (au minimum un facteur 10).

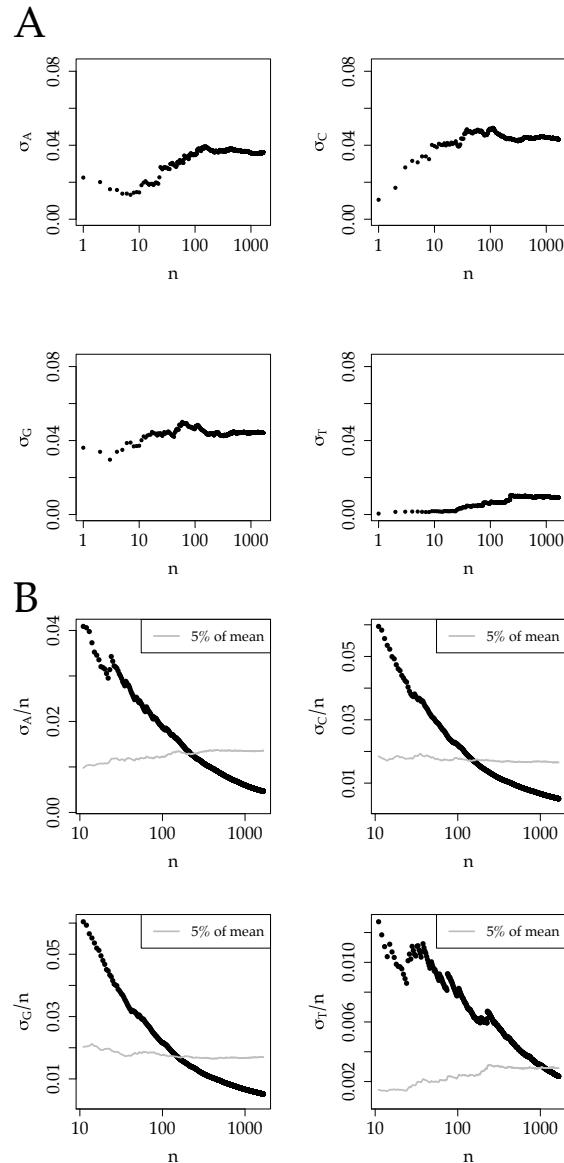


FIGURE 3.4 – Estimation de la convergence.

(A) Écart-type  $\sigma_b$  de la distribution  $w_{i,b}$  estimé à partir de  $n$  échantillons indépendants. Ces échantillons sont pris toutes les 30 itérations au sein de la chaîne MCMC, soit environ 3 fois le temps de décorrélation. On observe qu'au bout de 100 itérations l'écart-type est stabilisé. Par ailleurs les fluctuations sont relativement faibles, au plus de l'ordre de  $\sim 10\%$  de la moyenne.

(B) La convergence est estimée grâce au Théorème Centrale Limite. L'écart-type de l'estimateur de la moyenne se comporte comme  $\sigma_b/\sqrt{n}$ . La précision demandée correspond à un écart-type  $\leq 5\%$  de la moyenne pour les 4 bases, ce qui correspond à l'intersection la plus tardive entre les courbes noire et grise (dans notre cas le cadran du bas à droite).

## 3.3. Calcul de la moyenne de la postérieure par une méthode MCMC

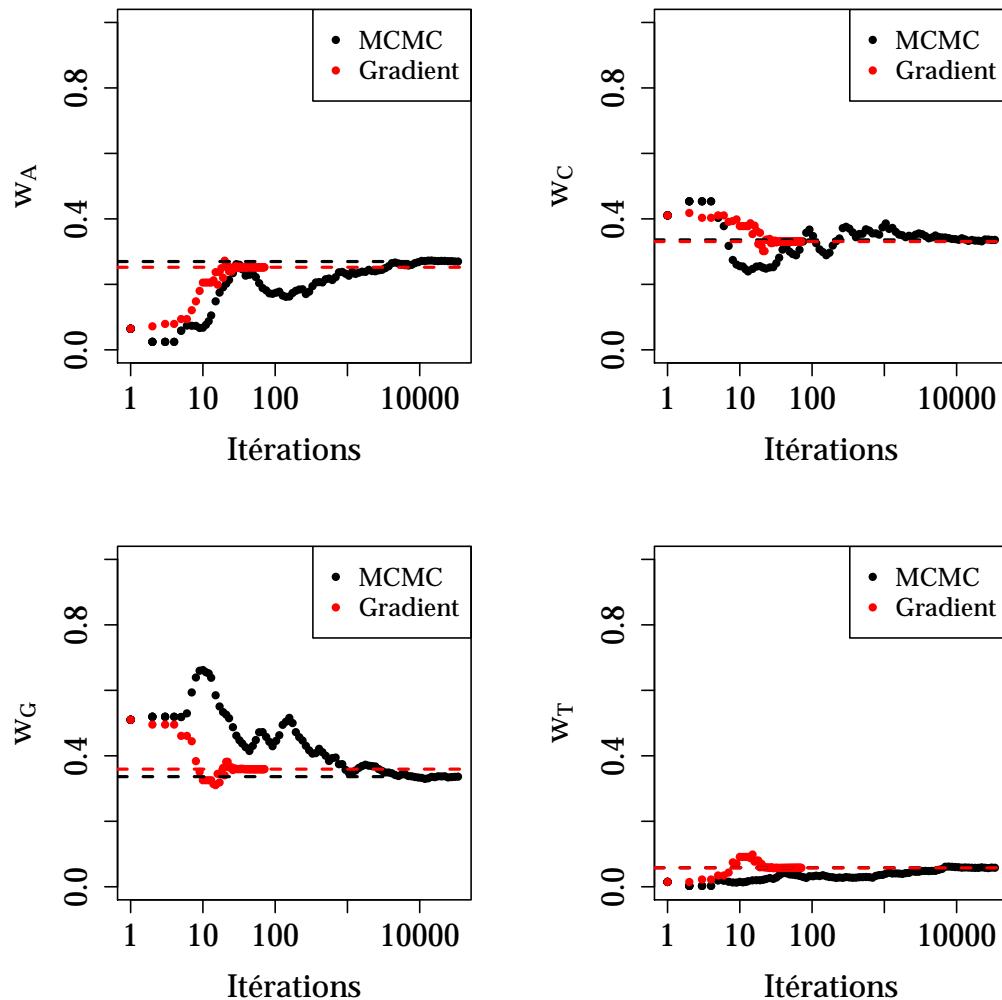


FIGURE 3.5 – Comparaison des approches par MCMC et par descente de gradient.

L'approche de maximisation de (l'opposé de) la postérieure modifiée par descente de gradient (rouge, cf article) est comparée à l'approche de calcul de la moyenne de la postérieure par MCMC (noir). Les deux méthodes sont initialisées au même  $w_i^{\text{inde}}$ . Alors que la méthode par descente de gradient converge rapidement ( $\sim 100$  itérations), l'approche MCMC converge plus lentement, dans ce cas plus de 10,000 itérations. Au final les deux approches convergent vers des quantités proches (lignes pointillées rouges et noires).

### 3.4 Conclusion et perspectives

Nous avons présenté Imogene, un algorithme bayésien utilisant la phylogénie de recherche de motifs et modules conduisant une régulation commune. Imogene est basé sur l'algorithme introduit par [Rouault et al. \(2010\)](#) dans le cas des Drosophiles, et l'étend au cas des mammifères. Nous avons présenté des tests d'Imogene sur des CRMs possédant une expression déterminée chez l'embryon de souris (tube neural et bourgeon de membre), et avons montré la capacité d'Imogene de prédire des CRMs conduisant à une expression similaire au sein de séquences intergéniques. Parmi les motifs générés par Imogene, certains sont associés à des régulateurs connus des étapes du développement considérées. Par ailleurs, nous avons montré que les motifs générés par Imogene pouvaient être utilisés pour générer un classifieur linéaire permettant d'associer un CRM donné à une classe liée à une expression spécifique. Ce classifieur montre des performances similaires à un classificateur basé sur des données biologiques spatio-temporelles extensives ([Zinzen et al., 2009](#)), mais ne nécessite que la connaissance de quelques séquences de chaque classe et fournit en plus la connaissance des motifs régulateurs.

Imogene peut être utilisé à partir de l'interface Mobyle de l'Institut Pasteur <http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::imogene>. Nous espérons ainsi qu'il pourra servir à des biologistes souhaitant mettre à jour des régulateurs putatifs dans des CRMs fonctionnels et détecter dans le génome des CRMs possédant une activité similaire.

---

## Chapitre 4

# Étude de la différenciation épidermale chez la drosophile

---

4.1 Concept d'optimum de Pareto . . . . .	109
4.2 Article . . . . .	111
4.3 Conclusion et perspectives . . . . .	112

## Introduction du chapitre 4

Nous présentons maintenant une application de Imogene au cas de la différenciation des trichomes (poils) chez la Drosophile, réalisée en collaboration avec l'équipe de Serge Plaza à l'Université Paul Sabatier. Plusieurs motifs ont été générés à partir de 14 CRMs connus pour réguler le processus de différenciation des trichomes. Parmi les motifs générés, deux d'entre eux ont montré une meilleure capacité à distinguer les CRMs positifs de l'ensemble d'apprentissage de CRMs négatifs. Le critère de distinction est basé sur l'optimisation de Pareto, caractérisant la satisfaction de plusieurs contraintes à la fois. Dans notre cas, les contraintes sont de maximiser le nombre de CRMs positifs trouvés par les motifs (maximisation de la sensibilité) tout en minimisant le nombre de faux positifs (maximisation de la spécificité). Ces critères définissent une frontière de Pareto de motifs optimum. En variant les différents paramètres de Imogene, nous avons trouvé deux motifs sur la frontière de Pareto. Parmi les deux motifs, l'un d'eux (« svbf<sub>7</sub> ») correspond au régulateur maître du processus de différenciation des trichomes, et l'autre (« blue motif ») est un motif nouveau. L'importance des deux motifs pour la régulation est montrée par mutagenèse. Par ailleurs, ces motifs permettent de distinguer des ChIP-seq pour *svb* liés à une régulation (dont le gène le plus proche subit une perte d'expression lors du KO de *svb*) des ChIP-seq ne l'étant pas. Ce travail montre donc un exemple de la possibilité d'utiliser Imogene sur un petit ensemble (ici 14 CRMs) de données biologiques fonctionnelles pour détecter des motifs nouveaux et fonctionnels.

## 4.1 Concept d'optimum de Pareto

Dans l'article qui suit, nous utilisons le principe d'optimum de Pareto, lié à la satisfaction simultanée de plusieurs contraintes. Le problème est le suivant. Étant donnés 14 CRMs positifs liés à une même régulation de la différenciation des trichomes chez l'embryon de Drosophile, et 25 CRMs négatifs ne conduisant aucune expression au stade de développement considéré, quels motifs permettent le mieux de distinguer les deux classes ? Le problème est similaire à celui de *pattern recognition* introduit dans l'article précédent en section 3.2. Néanmoins, nous avons ici adopté une démarche légèrement différente.

Des motifs sont appris sur les CRMs positifs pour différents seuils  $S_g$  (variant entre 7 et 13 bits). Ces motifs sont ensuite utilisés pour prédire les CRMs positifs parmi les CRMs initiaux. Un CRM est déclaré comme positif s'il contient au moins  $n$  sites conservés ( $n = 1, 2$ , ou  $3$ ) au-dessus d'un seuil  $S_s$  (variant entre 7 et 13 bits). Pour des paramètres  $S_g$ ,  $n$  et  $S_s$  donnés, un motif détectera un nombre FP de Faux Positifs (les CRMs négatifs qui sont prédits positifs) et omettra un nombre FN de Faux Négatifs (les CRMs positifs qui ne sont pas prédits comme tels). L'optimisation de Pareto consiste à trouver les paramètres  $S_g$ ,  $n$  et  $S_s$  qui minimisent à la fois FP et FN. Il est possible de minimiser différentes fonctions de coût pour FP et FN, attribuant des poids plus ou moins importants à l'une ou l'autre des contraintes. Dans notre cas, nous avons défini l'optimum de Pareto comme minimisant la fonction  $|FN + FP|$ . La droite  $|FN + FP| = cte$  contenant les meilleurs optima de Pareto pour les différents motifs est la frontière de Pareto.

Dans notre cas, deux des motifs générés sont sur la frontière de Pareto : le motif svbF7 et le motif bleu (fig. 4.1), pour  $S_g = 10$  bits,  $n = 1$ ,  $S_s = 8.5$  bits. Ce sont par ailleurs les 2 premiers motifs générés à ce seuil. Les 3 motifs suivants sont montrés en gris. Le motif svbF7 correspond au TF *svb* (*Shavenbaby*), régulateur maître de la différenciation des trichomes. Le motif OvoQ6 correspond à la PWM de Transfac pour *svb*, et n'apporte pas une aussi bonne classification que svbF7. Le motif bleu, quant à lui, n'est pas connu. Nous montrons aussi le motif jaune, introduit dans l'article, qui n'est pas prédit par Imogene mais correspond à un motif de régulation ultra-conservé dans les de différentes espèces de drosophiles (Elemento and Tavazoie, 2005), et possédant un rôle fonctionnel investigué par mutagenèse.

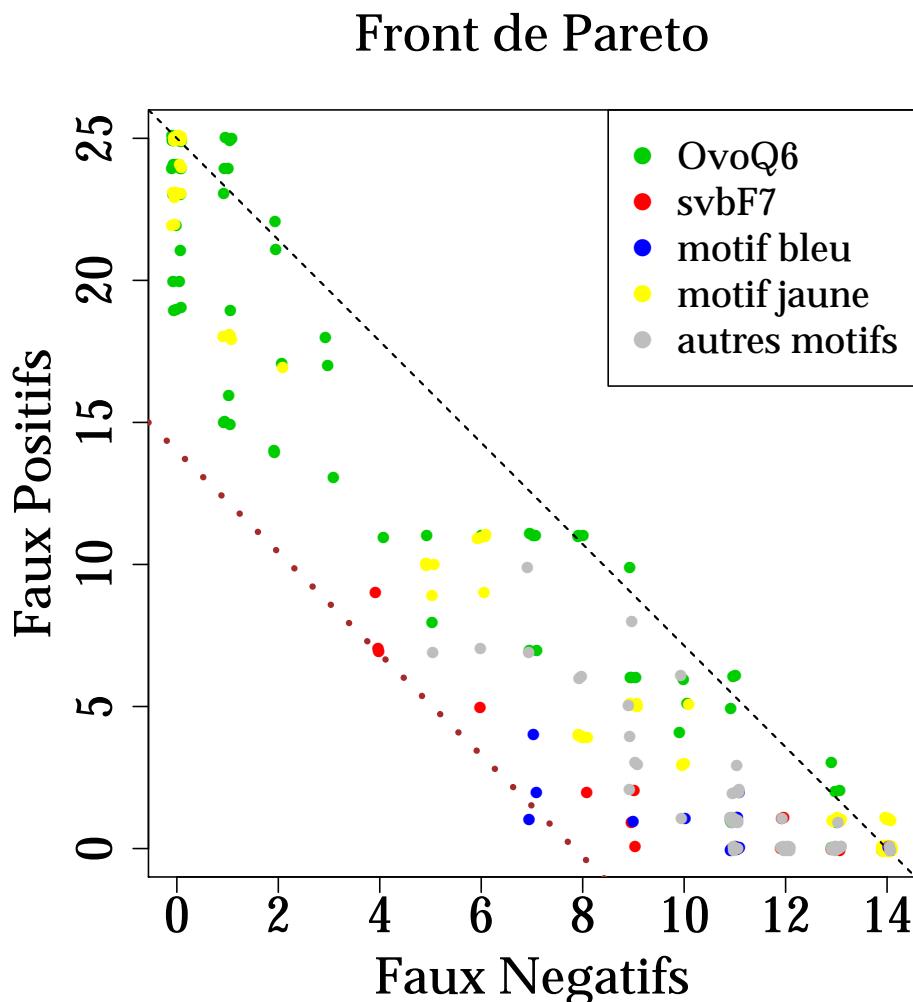


FIGURE 4.1 – Frontière d'efficacité de Pareto.

Illustration de l'optimisation de Pareto permettant la sélection des paramètres lors de la génération de motifs. Ici, nous voulons à la fois minimiser le nombre de Faux Positifs (FP) et le nombre de Faux négatifs (FN). Différents motifs sont successivement utilisés pour classer les séquences comme positives ou négatives, une séquence étant déclarée comme positive pour un motif si elle contient au moins  $n$  sites conservés au-dessus d'un seuil  $S_s$ . Les différents points correspondent aux différentes performances. Un léger bruit a été ajouté aux points pour rendre visibles ceux qui sont superposés (fonction jitter de R, paramètre factor= 0.5). Le motif svbF7 et le motif bleu sont tous deux optimaux pour des paramètres bien définis  $S_g = 10$  bits,  $n = 1$ ,  $S_s = 8.5$  bits : ils dessinent la frontière de Pareto  $|FP + FN| = cte$  en pointillés bruns.

## **4.2 Article**

### 4.3 Conclusion et perspectives

Imogene a été appliqué sur un ensemble d'apprentissage composé de 14 CRMs régulant la différenciation des trichomes chez l'embryon de Drosophile. Les différents paramètres (seuils de génération, de détection) ont été optimisés par une approche de Pareto visant à maximiser le nombre de CRMs positifs prédis par les motifs tout en minimisant le nombre de CRMs négatifs prédis parmi un ensemble de 25 CRMs n'ayant aucune activité au stade de développement considéré. Deux motifs ont été trouvé par cette approche : le motif svbF7 correspondant à *svb*, le régulateur maître de la différenciation des trichomes, et un nouveau motif, le motif bleu, que nous n'avons pas pu associer à un motif connu.

La validité de ces motifs a été montrée par mutagenèse (figures 3, 4 et 5 de l'article). Par ailleurs, ces motifs sont prédictifs des ChIP-seq de *svb* fonctionnels, c'est-à-dire ceux qui sont associés à un gène dont l'expression diminue chez les mutants *svb* (figure S5). Les CRMs fonctionnels possèdent une grande variété de grammaire des sites de fixation (figures 5 et 7 de l'article), un résultat similaire à celui obtenu par [Zinzen et al. \(2009\)](#) dans le cas de la différenciation de différents tissus chez l'embryon de drosophile. Plusieurs entrées (combinaisons de TFs) mènent à une sortie (motif d'expression du gène rapporteur) similaire : cette flexibilité est réminiscente du modèle *billboard* introduit en section [1.5.2](#). Néanmoins, bien qu'il soit clair que la grammaire des sites soit différente entre différents CRMs, cette grammaire semble relativement bien conservée au cours de l'évolution d'un CRM (figures 8, S8A, et S8B). Enfin, dans la plupart des cas on observe l'absence de *clustering* de motifs homotypiques sur les CRMs, bien que les motifs soient présents en plus grand nombre dans les loci des gènes régulés par rapport à des gènes non régulés (figure S2). Une explication possible est que l'abondance de sites dans l'environnement du CRM permet d'augmenter localement la concentration du TF pour faciliter son recrutement *in vivo* au niveau des CRMs possédant des sites de forte affinité.

Il serait à présent intéressant de caractériser plus en avant le motif bleu généré par l'approche *de novo*. Une possibilité serait d'utiliser la technique de simple hybride présentée en [1.4.2](#) afin d'identifier la protéine associé au motif bleu, en utilisant comme appât les protéines connues de la drosophile et comme proie le site consensus du motif bleu.

---

## Chapitre 5

# Étude de la différenciation musculaire chez la souris

---

5.1 Obtention d'une PWM pour MEF3 . . . . .	115
5.2 Article . . . . .	118

## Introduction du chapitre 5

Cette dernière partie est consacrée à l'étude de la différenciation musculaire. Ce travail a été effectué en collaboration avec l'équipe de Pascal Maire à l'Institut Cochin, qui m'a accueilli et m'a permis de réaliser une partie des expériences présentées, avec l'aide de Iori Sakakibara. Nous nous intéressons ici particulièrement aux homéoprotéines Six1 et Six4 (référées dans la suite par Six1,4) qui sont des facteurs de transcription impliqués dans la régulation des stades successifs de la myogenèse. Ils activent en effet les TFs nécessaires à l'engagement de cellules pluripotentes dans la voie myogénique et à leur différenciation : Pax3 ainsi que les Facteurs de Régulation Myogénique Myf5, Mrf4, Myod et Myog. Cette régulation passe notamment par la fixation des protéines Six sur un site MEF3 d'environ 10bp sur les séquences régulatrices du gène dont l'expression sera augmentée.

Nous nous sommes d'abord intéressé aux cibles transcriptionnelles de Six1,4 à l'échelle du génome, ainsi qu'à leurs corégulateurs. Pour cela, nous avons d'abord réalisé un modèle PWM des sites MEF3, que nous avons utilisé pour la prédiction de sites. Nous avons par ailleurs regroupé sur UCSC de nombreuses données relatives à la différenciation musculaire : ChIP-seq de TFs et des marques épigénétiques des histones, données d'expression de type RNAseq, sites de fixation conservés que nous avons obtenus par analyse bioinformatique, etc. Cet ensemble de données permet à lui seul d'envisager le contexte de régulation de certains gènes d'intérêt, et nous présentons plusieurs validations de prédictions obtenues par ces analyses bioinformatiques.

Par ailleurs, nous nous sommes intéressé à l'action concertée ou *synergie* entre Six1,4 et MyoD au cours de la différenciation musculaire. Nous avons trouvé un certain nombre de CRMs fixés par MyoD et possédant un site MEF3 conservé chez les mammifères, et dont le gène le plus proche n'est activé qu'en présence de MyoD et de Six1,4. Parmi les gènes en question figurent Myog, TF requis pour la différenciation terminale, et des gènes structuraux comme *Tnncl* ou *Ttn*. Nous avons testé l'activité des CRMs prédis et en avons trouvé 72% qui récapitulent l'expression du gène le plus proche lorsqu'ils sont testés par transfection avec un rapporteur Luciférase. Nous avons par ailleurs utilisé Imogene pour prédire des corégulateurs et avons testé l'importance des sites par mutagenèse.

## 5.1 Obtention d'une PWM pour MEF3

La première étape de notre collaboration avec l'équipe de P. Maire a été de construire un motif pour les sites de fixation MEF3 associés aux protéines Six1,4 chez la souris. Pour ce faire, nous avons utilisé le fait qu'un certain nombre de sites de fixation avaient déjà été testés par retard sur gel (EMSA ou *Electrophoretic Mobility Shift Assay*), que ce soit dans la littérature ou au sein du laboratoire de P. Maire. Ces expériences ont permis de définir 32 sites positifs (table 5.1) et 14 sites négatifs (table 5.2).

myog-spitz-pascal	GAAACCTGA	aldo-spitz-pascal	GAAACCTGA
pax3-pascal	GAAATCTAA	ato-pascal	GTCATTTGA
myf5-pascal	GTAACTGGA	kcnel-niro	GATAACCGA
myod-DRR-lhonore-pascal	GAAACCGGA	myf5-personal-josiane	GAAATTAA
Mlc-hox-pascal	GTAATTAA	pax3-1-personal-josiane	GAAATGTAA
atp2a1-1-niro-pascal	GTAACTGGA	pax3-2-personal-josiane	GTTACTGGA
atp2a1-2-niro	GTAACCTGA	pvalb-personal-iori	GTAACCTGA
tnnc2-niro	GAAATTAA	Utrophine-pascal	GTCACCTGA
Nrk-pascal	GCAAGGCGA	Na-K-ATPase-pascal	GCAACCTGA
sarcalumenin-pascal	GGAACTTGA	myf5-2-pascal	GCAACCTGA
myl1-2-niro	GAAATTGAA	myf5-sat-pascal	GAAATCTGA
ifitm3-niro	GTAATTGA	Ibx1-pascal	GCCACCTGA
mybph-2-niro	GAAATCTGA	MCK-pascal	GACACCCGA
myeov2-niro	GAAACTTGA	IgfBp5-pascal	GCAATTGAA
		troponine-C-low-pascal	GTAACCTGA
		sarcalumenin-pascal	GGAACTTGA
		Nrk-pascal	GCAAGGCGA
		Tp4-pascal	GCAAGCAGA

TABLE 5.1 – Sites MEF3 positifs.

Le tableau de gauche correspond aux 14 sites utilisés pour l'apprentissage (avec conservation chez mammifères), et la partie de droite aux sites utilisés pour la courbe ROC de la figure 5.1.

Il est à noter qu'ici les sites négatifs n'ont pas été choisis au hasard mais en utilisant leur ressemblance avec le site consensus de MEF3 sur le promoteur de *Myog*. Les sites positifs ont

myf5-1-personal	GACAGTGGA
myf5-2-personal	GTAACCTCA
myf5-mut	GTAACTGGG
pax3-1-pascal	GGAACTTGA
pax3-2-pascal	GTATTAATA
pax3-3-pascal	GGATAAAGA
pax3-4-pascal	GCTAATTGA
pax3-5-pascal	GAAAGATTA
pax3-6-pascal	GCTCTCTGA
pax3-7-pascal	GAGCCCTGA
pvalb-1-personal	GCACAATGA
pvalb-2-personal	GCAGGCTGA
unknown-1	GCAATCTGA
unknown-2	GAGTCCTGA

TABLE 5.2 – Sites MEF3 négatifs.

Sites négatifs utilisés pour la courbe ROC de la figure 5.1.

été partagés en un ensemble d'apprentissage composé de 14 séquences pour lesquelles les séquences orthologues chez les autres mammifères étaient disponibles, et en un ensemble test de 18 sites.

Nous avons alors cherché la PWM apprise sur l'ensemble d'apprentissage qui distinguait le mieux les sites positifs de l'ensemble test des sites négatifs. Nous avons pour cela généré deux types de PWM : des PWMs « référence » apprises sur les 14 séquences d'apprentissage, et des PWMs « avec évolution » apprises avec Imogene à partir des alignements des séquences d'apprentissage avec leurs orthologues. Le seuil de génération a été varié entre 7 et 12 bits, et les deux modèles d'évolution (*Felsenstein* et *Halpern-Bruno*) ont été utilisés. Pour chaque PWM, une courbe ROC (pour *Receiver Operating Characteristic*) peut être construite indiquant pour différents seuils de détection  $S_s$  la proportion de séquences positives détectées dans l'ensemble test (Vrais Positifs) et la proportion de séquences négatives détectées (Faux Positifs). Nous montrons en figure 5.1 les meilleures courbes ROC obtenues pour le cas référence et le cas avec évolution. Les deux courbes sont relativement similaires et permettent de détecter  $\sim 45\%$

## 5.1. Obtention d'une PWM pour MEF3

des sites positifs sans détecter aucun négatif pour un seuil de l'ordre de  $S_g$ . La PWM avec évolution, nommée MEF3evolh6, montre une légère amélioration du signal à haut seuil (petits Faux Positifs) par rapport à la PWM référence, et nous avons donc utilisé celle-ci dans la suite de notre travail.

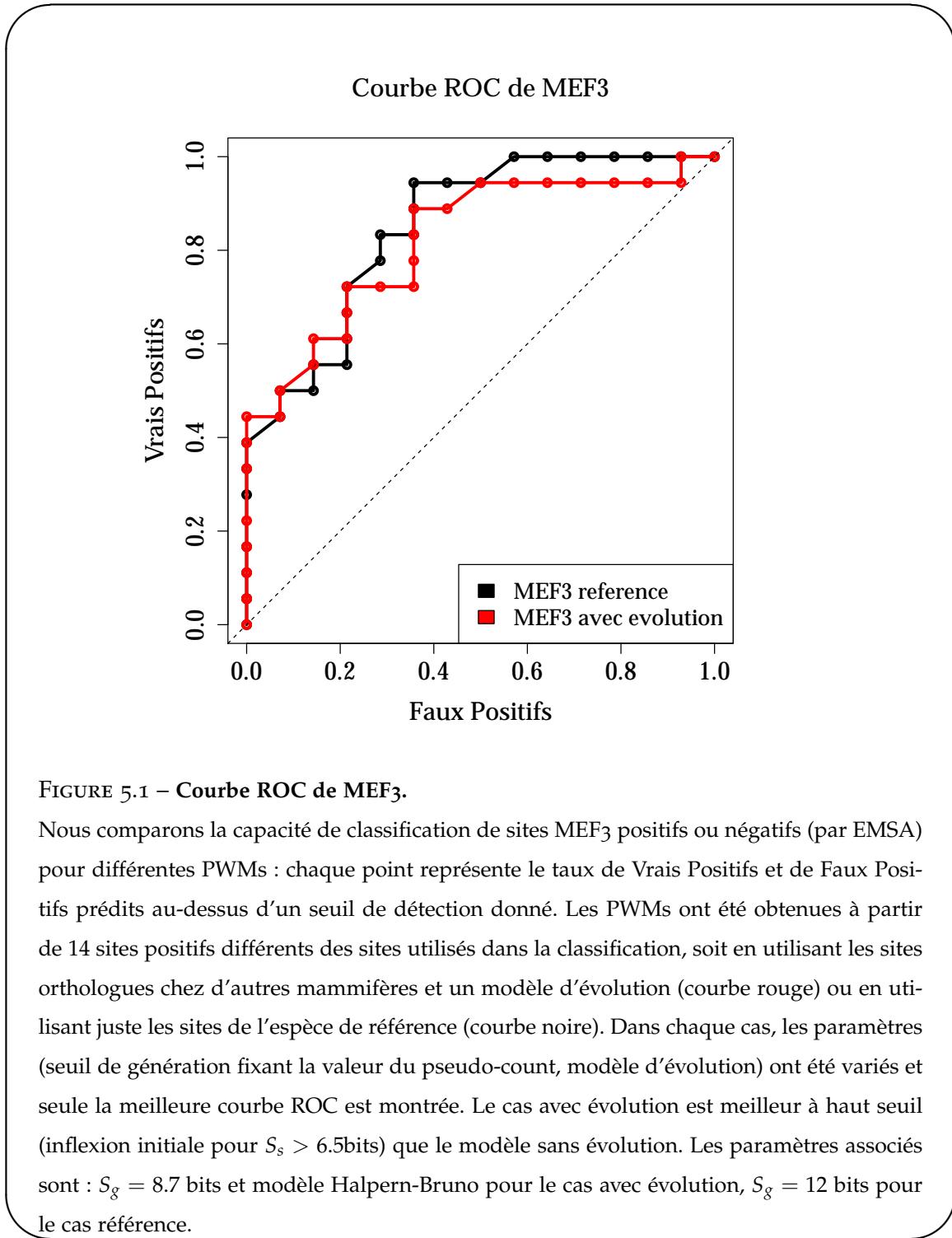


FIGURE 5.1 – Courbe ROC de MEF3.

Nous comparons la capacité de classification de sites MEF3 positifs ou négatifs (par EMSA) pour différentes PWMs : chaque point représente le taux de Vrais Positifs et de Faux Positifs prédits au-dessus d'un seuil de détection donné. Les PWMs ont été obtenues à partir de 14 sites positifs différents des sites utilisés dans la classification, soit en utilisant les sites orthologues chez d'autres mammifères et un modèle d'évolution (courbe rouge) ou en utilisant juste les sites de l'espèce de référence (courbe noire). Dans chaque cas, les paramètres (seuil de génération fixant la valeur du pseudo-count, modèle d'évolution) ont été variés et seule la meilleure courbe ROC est montrée. Le cas avec évolution est meilleur à haut seuil (inflexion initiale pour  $S_g > 6.5$ bits) que le modèle sans évolution. Les paramètres associés sont :  $S_g = 8.7$  bits et modèle Halpern-Bruno pour le cas avec évolution,  $S_g = 12$  bits pour le cas référence.

## 5.2 Article

## **Conclusion du chapitre 5**



---

## Annexe A

# Statistiques génomiques

---

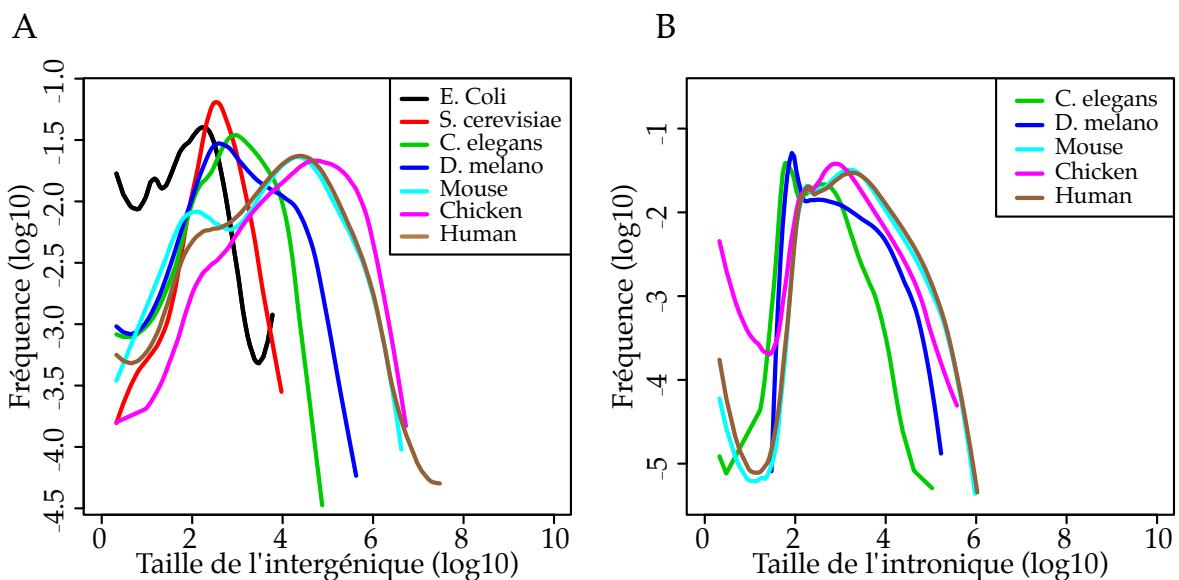
Nous nous intéressons ici aux distributions de tailles intergéniques et introniques dans les génomes de différentes espèces, et aux questions qu'elles soulèvent.

Nous l'avons vu en introduction (section [1.7.1](#)), l'interface Galaxy permet d'obtenir un certain nombre d'annotations génomiques à partir de différentes bases de données, comme UCSC. Les annotations génétiques consistent généralement en des coordonnées sur le génome de TSSs et de leurs exons et introns. Ceux-ci sont associés à un gène, qui peut avoir plusieurs TSSs différents et dont le transcript peut avoir plusieurs épissages alternatifs.

Nous avons utilisé Galaxy pour récupérer les annotations génomiques de différentes espèces, allant de l'unicellulaire à l'homme : la bactérie *Escherichia coli*, la levure *Saccharomyces cerevisiae*, le ver *Caenorhabditis elegans*, la mouche *Drosophila melanogaster*, la souris *Mus Musculus*, le poulet *Gallus gallus* et l'homme *Homo Sapiens*.

Les données intergéniques ont été obtenues en prenant les coordonnées complémentaires à l'ensemble fusionné des transcripts annotés. Dit autrement, chaque gène a été défini par les coordonnées les plus extrêmes de ses transcripts alternatifs, et les régions entre deux gènes définissent les régions intergéniques. La distribution de la taille de l'intergénique pour ces différentes espèces est montrée en figure [A.1A](#). On observe une loi proche d'une log-normale et qui semble, à une remise à l'échelle près, être relativement conservée chez les différentes espèces. Les distributions sont essentiellement unimodales, mais on note l'apparition d'un deuxième pic à  $\sim 100\text{bp}$  chez la souris et l'humain, peut-être dû à l'annotation récente de nombreuses régions transcrivées non codantes proches des gènes. On note par ailleurs l'inflation importante de la taille des régions intergéniques entre la bactérie (180 bp en moyenne, médiane à 100bp) et l'homme (80kb de moyenne, médiane à 15kb).

## Annexe A. Statistiques génomiques



**FIGURE A.1 – Distribution des tailles intergéniques et introniques chez différentes espèces.**

Distributions log-log de la taille des régions intergéniques (A) et introniques (B) chez différentes espèces. Les histogrammes sont réalisés avec un intervalle de 0.05 (en log 10) puis lissés avec l'estimateur local LOESS de paramètre  $span = 0.3$  (logiciel R). (A) Les régions intergéniques sont définies comme les régions complémentaires aux régions transcrtes (données UCSC), celles-ci étant préalablement fusionnées pour éviter les redondances liées aux multiples transcrits d'un même gène. De la bactérie à l'homme, on observe une inflation de la quantité de génome non codant. (B) Les régions introniques sont définies par le fait qu'elles sont entourées par deux exons d'un même gène. Pour pouvoir être épissées lors de la maturation des preARNm, les introns doivent posséder des sites d'épissage, imposant une borne inférieure à leur taille pour que l'ARNm final soit fonctionnel.

Les données introniques montrées en figure A.1B exhibent quant à elles un comportement très stéréotypé. Toutes les distributions présentent un pic fort autour de 80bp, que l'on peut interpréter comme une taille minimale pour que l'épissage puisse avoir lieu, et l'apparition d'un deuxième pic à 1kb chez les vertébrés. On observe par ailleurs que la présence d'introns longs semble corrélée à la taille des génomes.

La présence de distributions des longueurs d'intergénique et d'intronique stéréotypées au sein d'organismes aussi divers que la bactérie, le ver ou l'homme laisse à penser qu'il existe des mécanismes universels régissant la croissance des régions non codantes du génome. Il serait donc intéressant de comprendre quels mécanismes d'insertions-délétions (indels) permettent de modéliser ce phénomène. De manière générale, l'équation maîtresse caractérisant l'évolution de la distribution  $P(L, t)$  des longueurs intergéniques  $L$  au cours du temps  $t$  s'écrit

$$P(L, t + dt) = P(L, t) + \int_0^{\infty} dL' [P(L', t)\tau(L' \rightarrow L) - P(L, t)\tau(L \rightarrow L')] dt \quad (\text{A.1})$$

où  $\tau(L \rightarrow L')$  est le taux de transition de la taille  $L'$  vers la taille  $L$ , qui dépend des insertions-délétions. On imagine un mécanisme similaire pour les introns, avec une longueur minimale  $L_{\min}$ . De nombreuses données sur les statistiques des indels ont récemment été rendues accessibles chez l'homme dans le cadre du projet 1000 genomes (Mills et al., 2011) ainsi que chez la souris (Yalcin et al., 2011). Ces données pourraient servir à décrire ces transitions et voir si l'on peut reproduire les lois observées.



---

# 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. (Pages [41](#) et [84](#).)
- Alon, U. (2007a). An Introduction to Systems Biology : Design Principles of Biological Circuits (Mathematical and Computational Biology Series vol 10). (Page [17](#).)
- Alon, U. (2007b). Network motifs : theory and experimental approaches. *Nat Rev Genet*, 8(6) :450–461. (Page [17](#).)
- Asp, P., Blum, R., Vethantham, V., Parisi, F., Micsinai, M., Cheng, J., Bowman, C., Kluger, Y., and Dynlacht, B. D. (2011). PNAS Plus : Genome-wide remodeling of the epigenetic landscape during myogenic differentiation. *Proceedings of the National Academy of Sciences*, pages 1–11. (Page [62](#).)
- Attanasio, C., Reymond, A., Humbert, R., Lyle, R., Kuehn, M. S., Neph, S., Sabo, P. J., Goldy, J., Weaver, M., Haydock, A., Lee, K., Dorschner, M., Dermitzakis, E. T., Antonarakis, S. E., and Stamatoyannopoulos, J. A. (2008). Assaying the regulatory potential of mammalian conserved non-coding sequences in human cells. *Genome Biol*, 9(12) :R168. (Page [53](#).)
- Aurell, E., d'Hérouël, A., Malmnäs, C., and Vergassola, M. (2007). Transcription factor concentrations versus binding site affinities in the yeast *S. cerevisiae*. *Physical biology*, 4 :134. (Page [27](#).)
- Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.-F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., and Bulyk, M. L. (2009). Diversity and Complexity in DNA Recognition by Transcription Factors. *Science*, 324(5935) :1720–1723. (Pages [67](#) et [71](#).)
- Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2 :28–36. (Page [85](#).)
- Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003). Modeling dependencies in protein-DNA binding sites. In *Proceedings of the seventh annual international conference on Research in*

## Bibliographie

---

- computational molecular biology*, RECOMB '03, pages 28–37. ACM, New York, NY, USA. ISBN 1-58113-635-8. (Pages [68](#), [70](#), [71](#) et [72](#).)
- Bartel, D. P. (2009). MicroRNAs : target recognition and regulatory functions. *Cell*, 136(2) :215–33. (Page [14](#).)
- Baxter, R. (2007). *Exactly Solved Models in Statistical Mechanics*. Dover Books on Physics Series. DOVER PUBN Incorporated. ISBN 9780486462714. (Page [77](#).)
- Baylies, M. K., Bate, M., and Ruiz Gomez, M. (1998). Myogenesis : a view from Drosophila. *Cell*, 93(6) :921–7. (Page [19](#).)
- Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98(3) :387–96. (Page [41](#).)
- Benos, P. V., Bulyk, M. L., and Stormo, G. D. (2002). Additivity in protein-DNA interactions : how good an approximation is it ? *Nucleic Acids Res*, 30(20) :4442–51. (Page [69](#).)
- 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 [21](#).)
- Berg, O. G., Winter, R. B., and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, 20(24) :6929–48. (Page [21](#).)
- Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., 3rd, and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11) :1429–35. (Page [29](#).)
- Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., and Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A*, 99(2) :757–62. (Page [50](#).)
- Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev*, 16(1) :6–21. (Page [13](#).)

- Blackwell, T. K. and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science*, 250(4984) :1104–10. (Page 30.)
- Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganière, J., Lefèvre, C., Deblois, G., Giguère, V., Ferretti, V., Bergeron, D., Coulombe, B., and Robert, F. (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res*, 16(5) :656–68. (Page 53.)
- Blau, H. M., Pavlath, G. K., Hardeman, E. C., Chiu, C. P., Silberstein, L., Webster, S. G., Miller, S. C., and Webster, C. (1985). Plasticity of the differentiated state. *Science*, 230(4727) :758–66. (Page 8.)
- Bolouri, H. and Davidson, E. H. (2002). Modeling DNA sequence-based cis-regulatory gene networks. *Dev Biol*, 246(1) :2–13. (Page 40.)
- Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H. H., and Liu, E. T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*, 18(11) :1752–1762. (Page 45.)
- Brazma, A., Parkinson, H., Schlitt, T., and Shojatalab, M. (2001). A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays. [http://www.ebi.ac.uk/microarray/biology\\_intro.html](http://www.ebi.ac.uk/microarray/biology_intro.html). (Page 5.)
- Buchler, N. E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9) :5136–41. (Page 40.)
- Bulyk, M. L., Johnson, P. L. F., and Church, G. M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5) :1255–61. (Page 67.)
- Bussemaker, H. J., Li, H., and Siggia, E. D. (2000). Building a dictionary for genomes : identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A*, 97(18) :10096–100. (Page 88.)
- Campbell, C. T. and Kim, G. (2007). SPR microscopy and its applications to high-throughput analyses of biomolecular binding events and their kinetics. *Biomaterials*, 28(15) :2380–92. (Page 29.)

## Bibliographie

---

- Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G. J., Parker, M. H., Macquarrie, K. L., Davison, J., Morgan, M. T., Ruzzo, W. L., Gentleman, R. C., and Tapscott, S. J. (2010a). Genome-wide MyoD binding in skeletal muscle cells : a potential for broad cellular reprogramming. *Developmental Cell*, 18(4) :662–74. (Pages [52](#) et [95](#).)
- Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G. J., Parker, M. H., Macquarrie, K. L., Davison, J., Morgan, M. T., Ruzzo, W. L., Gentleman, R. C., and Tapscott, S. J. (2010b). Genome-wide MyoD binding in skeletal muscle cells : a potential for broad cellular reprogramming. *Developmental Cell*, 18(4) :662–74. (Page [62](#).)
- Carlson, C. D., Warren, C. L., Hauschild, K. E., Ozers, M. S., Qadir, N., Bhimsaria, D., Lee, Y., Cerrina, F., and Ansari, A. Z. (2010). Specificity landscapes of DNA binding molecules elucidate biological function. *Proc Natl Acad Sci U S A*, 107(10) :4544–9. (Page [30](#).)
- Carninci, P., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6) :626–35. (Page [39](#).)
- Carvajal, J. J., Keith, A., and Rigby, P. W. J. (2008). Global transcriptional regulation of the locus encoding the skeletal muscle determination genes Mrf4 and Myf5. *Genes & development*, 22(2) :265–76. (Pages [39](#) et [40](#).)
- Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3) :167–174. (Page [69](#).)
- Chan, B. Y. and Kibler, D. (2005). Using hexamers to predict cis-regulatory motifs in *Drosophila*. *BMC Bioinformatics*, 6 :262. (Page [95](#).)
- Cheng, Y., King, D. C., Dore, L. C., Zhang, X., Zhou, Y., Zhang, Y., Dorman, C., Abebe, D., Kumar, S. A., Chiaromonte, F., Miller, W., Green, R. D., Weiss, M. J., and Hardison, R. C. (2008). Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif. *Genome Res*, 18(12) :1896–905. (Page [55](#).)
- Chung, J. H., Whiteley, M., and Felsenfeld, G. (1993). A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in *Drosophila*. *Cell*, 74(3) :505–14. (Page [41](#).)

- Cordaux, R. and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10) :691–703. (Page 47.)
- Davis, R. L., Weintraub, H., and Lassar, A. B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell*, 51(6) :987–1000. (Page 9.)
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38. (Page 87.)
- Dermitzakis, E. T. and Clark, A. G. (2002). Evolution of transcription factor binding sites in Mammalian gene regulatory regions : conservation and turnover. *Mol Biol Evol*, 19(7) :1114–21. (Page 45.)
- Dermitzakis, E. T., Reymond, A., and Antonarakis, S. E. (2005). Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet*, 6(2) :151–7. (Page 53.)
- D'haeseleer, P. (2006). How does DNA sequence motif discovery work ? *Nat Biotechnol*, 24(8) :959–61. (Page 85.)
- Djordjevic, M., Sengupta, A. M., and Shraiman, B. I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Res*, 13(11) :2381–90. (Page 24.)
- Donaldson, I. J., Chapman, M., Kinston, S., Landry, J. R., Knezevic, K., Piltz, S., Buckley, N., Green, A. R., and Göttgens, B. (2005). Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. *Hum Mol Genet*, 14(5) :595–601. (Page 54.)
- Elemento, O. and Tavazoie, S. (2005). Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol*, 6(2) :R18. (Page 109.)
- ENCODE Project Consortium, et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414) :57–74. (Page 64.)
- Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shores, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E.

## Bibliographie

---

- (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345) :43–9. (Page 55.)
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences : a maximum likelihood approach. *J Mol Evol*, 17(6) :368–76. (Pages 89, 90 et 91.)
- Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*, 9(5) :397–405. (Page 47.)
- Fields, D. S., He, Y., Al-Uzri, A. Y., and Stormo, G. D. (1997). Quantitative specificity of the Mnt repressor. *J Mol Biol*, 271(2) :178–94. (Page 30.)
- Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., and Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, pages 1–5. (Page 48.)
- Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29(2-3) :131–163. (Page 70.)
- Furusawa, C. and Kaneko, K. (2012). A Dynamical-Systems View of Stem Cell Biology. *Science*, 338(6104) :215–217. (Page 6.)
- Gerland, U., Moroz, J., and Hwa, T. (2002). Physical constraints and functional characteristics of transcription factor–DNA interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 99(19) :12015. (Pages 21, 24, 25 et 26.)
- Giocomo, L. M., Moser, M.-B., and Moser, E. I. (2011). Computational models of grid cells. *Neuron*, 71(4) :589–603. (Page 23.)
- Grad, Y. H., Roth, F. P., Halfon, M. S., and Church, G. M. (2004). Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D.pseudoobscura*. *Bioinformatics*, 20(16) :2738–50. (Page 93.)
- Graf, T. and Enver, T. (2009). Forcing cells to change lineages. *Nature*, 462(7273) :587–94. (Page 9.)
- Greer, E. L. and Shi, Y. (2012). Histone methylation : a dynamic mark in health, disease and inheritance. *Nat Rev Genet*, 13(5) :343–57. (Page 14.)

- Grendar Jr, M. and Grendar, M. (2001). MiniMax Entropy and Maximum Likelihood : complementarity of tasks, identity of solutions. In *AIP Conference Proceedings*, volume 568, page 49. (Page 75.)
- Gurdon, J. B. and Melton, D. A. (2008). Nuclear reprogramming in cells. *Science*, 322(5909) :1811–5. (Page 9.)
- Halpern, A. and Bruno, W. (1998). Evolutionary distances for protein-coding sequences : modeling site-specific residue frequencies. *Molecular biology and evolution*, 15(7) :910. (Page 92.)
- Hammond, S. M., Caudy, A. A., and Hannon, G. J. (2001). Post-transcriptional gene silencing by double-stranded RNA. *Nat Rev Genet*, 2(2) :110–9. (Page 14.)
- Hannon, G. J. (2002). RNA interference. *Nature*, 418(6894) :244–51. (Page 14.)
- Hardison, R. C. and Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics*, 13(7) :469–483. (Pages 38 et 55.)
- Hartwell, L., Hopfield, J., Leibler, S., and Murray, A. (1999). From molecular to modular cell biology. *Nature*, 402(6761) :47. (Pages 51 et 56.)
- Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2) :160–74. (Page 91.)
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109. (Page 97.)
- Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, 39(3) :311–8. (Page 55.)
- Heintzman, N. D., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243) :108–12. (Pages 40 et 55.)
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 38(4) :576–89. (Page 95.)

## Bibliographie

---

- Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods*, 6(4) :283–9. (Page [36](#).)
- Hong, J.-W., Hendrix, D. A., and Levine, M. S. (2008). Shadow enhancers as a source of evolutionary novelty. *Science*, 321(5894) :1314. (Page [47](#).)
- Hu, M., Yu, J., Taylor, J. M., Chinnaiyan, A. M., and Qin, Z. S. (2010). On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic acids research*, 38(7) :2154–2167. (Page [69](#).)
- Ivan, A., Halfon, M. S., and Sinha, S. (2008). Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs. *Genome Biol*, 9(1) :R22. (Page [94](#).)
- Jaynes, E. T. (1978). Where do we stand on maximum entropy ? pages 1–105. (Page [72](#).)
- Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9) :939–952. (Page [75](#).)
- Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E., and Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*, 20(6) :861–73. (Page [31](#).)
- Jolma, A., Yan, J., 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. (Pages [31](#), [67](#), [71](#) et [82](#).)
- Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J., and Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314) :430–435. (Page [48](#).)
- Kantorovitz, M. R., Kazemian, M., Kinston, S., Miranda-Saavedra, D., Zhu, Q., Robinson, G. E., Göttgens, B., Halfon, M. S., and Sinha, S. (2009). Motif-blind, genome-wide discovery of cis-

- regulatory modules in *Drosophila* and mouse. *Developmental Cell*, 17(4) :568–79. (Pages 52, 54, 93, 94 et 95.)
- Kantorovitz, M. R., Robinson, G. E., and Sinha, S. (2007). A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, 23(13) :i249–55. (Page 94.)
- Kaufmann, S. (1993). The origins of order. (Page 7.)
- Kazemian, M., Zhu, Q., Halfon, M. S., and Sinha, S. (2011). Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison. *Nucleic Acids Res*, 39(22) :9463–72. (Page 94.)
- 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 5.)
- Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res*, 17(12) :1919–31. (Page 53.)
- Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics*, 47(6) :713. (Page 92.)
- Kinney, J. B., Tkacik, G., and Callan, C. G. (2007). Precise physical models of protein-DNA interaction from high-throughput data. *Proc Natl Acad Sci USA*, 104(2) :501–6. (Page 29.)
- Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., and Kageyama, Y. (2010). Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science*, 329(5989) :336–9. (Page 14.)
- Krauth, W. (2006). *Statistical mechanics : algorithms and computations*, volume 13. Oxford University Press. (Pages 69, 96 et 99.)
- Kulessa, H., Frampton, J., and Graf, T. (1995). GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes Dev*, 9(10) :1250–62. (Page 9.)
- Kulkarni, M. M. and Arnosti, D. N. (2003). Information display by transcriptional enhancers. *Development*, 130(26) :6569–75. (Pages 42 et 43.)

## Bibliographie

---

- Lander, E. S., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921. (Page [59](#).)
- Lässig, M. (2007). From biophysics to evolutionary genetics : statistical aspects of gene regulation. *BMC Bioinformatics*, 8(Suppl 6) :S7. (Pages [20](#) et [25](#).)
- Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7(1) :41–51. (Pages [85](#) et [86](#).)
- Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., and Simon, I. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594) :799. (Pages [16](#) et [17](#).)
- Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*, 12(14) :1725–35. (Pages [39](#), [57](#) et [58](#).)
- 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. (Pages [44](#) et [45](#).)
- Liu, Y., Chu, A., Chakroun, I., Islam, U., and Blais, A. (2010). Cooperation between myogenic regulatory factors and SIX family transcription factors is important for myoblast differentiation (SI). *Nucleic acids research*. (Page [62](#).)
- Liu, Y.-H., Jakobsen, J. S., Valentin, G., Amarantos, I., Gilmour, D. T., and Furlong, E. E. M. (2009). A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development. *Developmental Cell*, 16(2) :280–91. (Page [18](#).)
- Loots, G. G. and Ovcharenko, I. (2004). rVISTA 2.0 : evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*, 32(Web Server issue) :W217–21. (Page [63](#).)
- Loots, G. G. and Ovcharenko, I. (2005). Dcode.org anthology of comparative genomic tools. *Nucleic Acids Res*, 33(Web Server issue) :W56–64. (Page [61](#).)

- Ludwig, M. Z., Bergman, C., Patel, N. H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403(6769) :564–7. (Page 45.)
- Maerkl, S. and Quake, S. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809) :233. (Page 28.)
- Majoros, W. H. and Ohler, U. (2010). Modeling the evolution of regulatory elements by simultaneous detection and alignment with phylogenetic pair HMMs. *PLoS Comput Biol*, 6(12) :e1001037. (Page 54.)
- Man, T. K. and Stormo, G. D. (2001). Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res*, 29(12) :2471–8. (Page 66.)
- Maniatis, T., Goodbourn, S., and Fischer, J. A. (1987). Regulation of inducible and tissue-specific gene expression. *Science*, 236(4806) :1237–45. (Page 39.)
- Masuya, H., Sezutsu, H., Sakuraba, Y., Sagai, T., Hosoya, M., Kaneda, H., Miura, I., Kobayashi, K., Sumiyama, K., Shimizu, A., Nagano, J., Yokoyama, H., Kaneko, S., Sakurai, N., Okagaki, Y., Noda, T., Wakana, S., Gondo, Y., and Shiroishi, T. (2007). A series of ENU-induced single-base substitutions in a long-range cis-element altering Sonic hedgehog expression in the developing mouse limb bud. *Genomics*, 89(2) :207–14. (Page 58.)
- McGregor, A., Orgogozo, V., Delon, I., Zanet, J., Srinivasan, D., Payre, F., and Stern, D. (2007). Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature*, 448(7153) :587–590. (Page 47.)
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21 :1087. (Page 97.)
- Mills, R. E., et al. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332) :59–65. (Page 123.)
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs : simple building blocks of complex networks. *Science*, 298(5594) :824–7. (Page 17.)

## Bibliographie

---

- Moses, A., Chiang, D., Pollard, D., Iyer, V., and Eisen, M. (2004). MONKEY : identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*, 5(12) :R98. (Page [90](#).)
- Moses, A. M., Chiang, D. Y., Kellis, M., Lander, E. S., and Eisen, M. B. (2003). Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol*, 3 :19. (Page [92](#).)
- Moses, A. M., Pollard, D. A., Nix, D. A., Iyer, V. N., Li, X.-Y., Biggin, M. D., and Eisen, M. B. (2006). Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol*, 2(10) :e130. (Page [45](#).)
- Nagaraj, V. H., O'flanagan, R. A., and Sengupta, A. M. (2008). Better estimation of protein-DNA interaction parameters improve prediction of functional sites. *BMC Biotechnol*, 8(1) :94. (Pages [30](#) et [31](#).)
- Nurse, P. and Hayles, J. (2011). The Cell in an Era of Systems Biology. *Cell*. (Page [10](#).)
- Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K., and Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*, 39(6) :730–2. (Page [45](#).)
- Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Rinaldi, N. J., Murray, H. L., Volkert, T. L., Schreiber, J., Rolfe, P. A., Gifford, D. K., Fraenkel, E., Bell, G. I., and Young, R. A. (2004). Control of pancreas and liver gene expression by HNF transcription factors. *Science*, 303(5662) :1378–81. (Pages [16](#) et [17](#).)
- Oliphant, A. R., Brandl, C. J., and Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides : analysis of yeast GCN4 protein. *Mol Cell Biol*, 9(7) :2944–9. (Page [30](#).)
- Ondek, B., Gloss, L., and Herr, W. (1988). The SV40 enhancer contains two distinct levels of organization. *Nature*, 333(6168) :40–5. (Page [39](#).)
- Panne, D. (2008). The enhanceosome. *Curr Opin Struct Biol*, 18(2) :236–42. (Page [43](#).)

- Park, P. J. (2009). ChIP-seq : advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10) :669–80. (Page 35.)
- Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B. L., Couronne, O., Eisen, M. B., Visel, A., and Rubin, E. M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118) :499–502. (Pages 43 et 53.)
- Perry, M. W., Boettiger, A. N., and Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the *Drosophila* embryo. *Proc Natl Acad Sci U S A*, 108(33) :13570–5. (Page 48.)
- Phillips, J. E. and Corces, V. G. (2009). CTCF : master weaver of the genome. *Cell*, 137(7) :1194–211. (Page 41.)
- Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E. M., Couronne, O., and Pennacchio, L. A. (2006). Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res*, 16(7) :855–63. (Page 43.)
- Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E. D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics*, 3 :30. (Pages 50 et 88.)
- Recillas-Targa, F., Pikaart, M. J., Burgess-Beusse, B., Bell, A. C., Litt, M. D., West, A. G., Gaszner, M., and Felsenfeld, G. (2002). Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proc Natl Acad Sci U S A*, 99(10) :6883–8. (Page 41.)
- Roh, T.-Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev*, 19(5) :542–52. (Page 55.)
- Rouault, H., Mazouni, K., Couturier, L., Hakim, V., and Schweisguth, F. (2010). Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny. *Proceedings of the National Academy of Sciences*, 107(33) :14615. (Pages 84, 95, 96 et 106.)

## Bibliographie

---

- Roulet, E., Busso, S., Camargo, A. A., Simpson, A. J. G., Mermod, N., and Bucher, P. (2002). High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol*, 20(8) :831–5. (Page 30.)
- Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development*, 132(4) :797–803. (Pages 57 et 58.)
- Sato, S., Ikeda, K., Shioi, G., Nakao, K., Yajima, H., and Kawakami, K. (2012). Regulation of Six1 expression by evolutionarily conserved enhancers in tetrapods. *Dev Biol*, 368(1) :95–108. (Page 63.)
- Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., and Gerstein, M. B. (2011). The real cost of sequencing : higher than you think ! *Genome Biol*, 12(8) :125. (Page 59.)
- Schirm, S., Jiricny, J., and Schaffner, W. (1987). The SV40 enhancer can be dissected into multiple segments, each with a different cell type specificity. *Genes Dev*, 1(1) :65–74. (Page 39.)
- Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, Â., Kutter, C., Brown, G. D., Marshall, A., Fliceck, P., and Odom, D. T. (2012). Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell*, 148(1-2) :335–348. (Page 47.)
- Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Fliceck, P., and Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding SI. *Science*, 328(5981) :1036–40. (Page 45.)
- Schoborg, T. A. and Labrador, M. (2010). The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. *J Mol Evol*, 70(1) :74–84. (Page 41.)
- Schones, D. E. and Zhao, K. (2008). Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet*, 9(3) :179–91. (Page 15.)
- Schroeder, M. D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E. D., and Gaul, U. (2004). Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol*, 2(9) :E271. (Page 50.)

- Sethna, J. (2006). *Statistical Mechanics : Entropy, Order Parameters and Complexity*. Oxford Master Series in Physics. OUP Oxford. ISBN 9780198566779. (Page 74.)
- Shen-Orr, S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*, 31(1) :64–68. (Page 17.)
- Shumaker-Parry, J. S., Aebersold, R., and Campbell, C. T. (2004). Parallel, quantitative measurement of protein binding to a 120-element double-stranded DNA array in real time using surface plasmon resonance microscopy. *Anal Chem*, 76(7) :2071–82. (Page 29.)
- Sinha, S. and He, X. (2007). MORPH : probabilistic alignment combined with hidden Markov models of cis-regulatory modules. *PLoS Comput Biol*, 3(11) :e216. (Page 54.)
- Sinha, S., Nimwegen, E. V., and Siggia, E. (2003). A probabilistic method to detect regulatory modules. *Bioinformatics*, 19(Suppl 1) :i292. (Pages 88 et 89.)
- Sinha, S., Schroeder, M. D., Unnerstall, U., Gaul, U., and Siggia, E. D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics*, 5 :129. (Page 53.)
- Sinha, S. and Tompa, M. (2000). A statistical method for finding transcription factor binding sites. *Proc Int Conf Intell Syst Mol Biol*, 8 :344–54. (Page 94.)
- Slutsky, M. and Mirny, L. A. (2004). Kinetics of protein-DNA interaction : facilitated target location in sequence-dependent potential. *Biophys J*, 87(6) :4021–35. (Page 21.)
- Smith, A. D., Sumazin, P., Xuan, Z., and Zhang, M. Q. (2006). DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci U S A*, 103(16) :6275–80. (Page 52.)
- Smith, A. D., Sumazin, P., and Zhang, M. Q. (2005). Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A*, 102(5) :1560–5. (Page 52.)
- Stormo, G. and Fields, D. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends in biochemical sciences*, 23(3) :109–113. (Page 22.)
- Stormo, G. D. and Zhao, Y. (2007). Putting numbers on the network connections. *Bioessays*, 29(8) :717–21. (Page 28.)

## Bibliographie

---

- Stormo, G. D. and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. *Nature Reviews Genetics*, 11(11) :751–60. (Pages 27 et 32.)
- Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4) :663–76. (Page 10.)
- Taylor, J., Tyekucheva, S., King, D. C., Hardison, R. C., Miller, W., and Chiaromonte, F. (2006). ESPERR : learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res*, 16(12) :1596–604. (Page 54.)
- Thurman, R. E., et al. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414) :75–82. (Page 37.)
- Tijssen, M. R., Cvejic, A., Joshi, A., Hannah, R. L., Ferreira, R., Forrai, A., Bellissimo, D. C., Oram, S. H., Smethurst, P. A., Wilson, N. K., Wang, X., Ottersbach, K., Stemple, D. L., Green, A. R., Ouwehand, W. H., and Göttgens, B. (2011). Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell*, 20(5) :597–609. (Page 55.)
- Tirosh, I., Weinberger, A., Bezalel, D., Kaganovich, M., and Barkai, N. (2008). On the relation between promoter divergence and gene expression evolution. *Mol Syst Biol*, 4 :159. (Page 45.)
- Trinklein, N. D., Aldred, S. J. F., Saldanha, A. J., and Myers, R. M. (2003). Identification and functional analysis of human transcriptional promoters. *Genome Res*, 13(2) :308–12. (Page 54.)
- Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment : RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968) :505–10. (Page 30.)
- U.S. Department of Energy (2001). Genomes to life : accelerating biological discovery (Office of Biological and Environmental Research and Office of Advanced Scientific Computing Research of the U.S. Department of Energy). [http://genomicscience.energy.gov/roadmap/GTLcomplete\\_web.pdf](http://genomicscience.energy.gov/roadmap/GTLcomplete_web.pdf). (Pages 11 et 12.)
- 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 13.)

- Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (2009a). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231) :854–8. (Page 55.)
- Visel, A., Rubin, E. M., and Pennacchio, L. A. (2009b). Genomic views of distant-acting enhancers. *Nature*, 461(7261) :199–205. (Pages 33 et 57.)
- Waddington, C. H. et al. (1957). The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.*, pages ix+–262. (Pages 6 et 7.)
- Wallace, J. A. and Felsenfeld, G. (2007). We gather together : insulators and genome organization. *Curr Opin Genet Dev*, 17(5) :400–7. (Page 41.)
- Wang, Q., Carroll, J. S., and Brown, M. (2005). Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol Cell*, 19(5) :631–42. (Page 40.)
- Warren, C. L., Kratochvil, N. C. S., Hauschild, K. E., Foister, S., Brezinski, M. L., Dervan, P. B., Phillips, G. N., Jr, and Ansari, A. Z. (2006). Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci U S A*, 103(4) :867–72. (Page 29.)
- Wasserman, W. and Fickett, J. (1998). Identification of regulatory regions which confer muscle-specific gene expression1. *Journal of molecular biology*, 278(1) :167–181. (Page 50.)
- Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4) :276–87. (Pages 23, 59 et 84.)
- Weintraub, H., Tapscott, S. J., Davis, R. L., Thayer, M. J., Adam, M. A., Lassar, A. B., and Miller, A. D. (1989). Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc Natl Acad Sci U S A*, 86(14) :5434–8. (Page 18.)
- Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2) :307–19. (Pages 48 et 49.)
- Wilczynski, B. and Furlong, E. E. M. (2010). Challenges for modeling global gene regulatory

## Bibliographie

---

- networks during development : Insights from Drosophila. *Developmental Biology*, 340(2) :161–169. (Page 40.)
- Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., Conboy, C. M., Vanes, L., Tybulewicz, V. L. J., Fisher, E. M. C., Tavaré, S., and Odom, D. T. (2008). Species-specific transcription in mice carrying human chromosome 21. *Science*, 322(5900) :434–8. (Page 45.)
- Wilson, M. D. and Odom, D. T. (2009). Evolution of transcriptional control in mammals. *Curr Opin Genet Dev*, 19(6) :579–85. (Pages 39, 45 et 46.)
- Winter, R. B., Berg, O. G., and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor-operator interaction : kinetic measurements and conclusions. *Biochemistry*, 20(24) :6961–77. (Page 21.)
- Winter, R. B. and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The Escherichia coli repressor-operator interaction : equilibrium measurements. *Biochemistry*, 20(24) :6948–60. (Page 21.)
- Wright, W. E., Binder, M., and Funk, W. (1991). Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. *Mol Cell Biol*, 11(8) :4104–10. (Page 30.)
- Xie, X., Lu, J., Kulkarni, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031) :338–45. (Page 53.)
- Yalcin, B., Wong, K., Agam, A., Goodson, M., Keane, T. M., Gan, X., Nellåker, C., Goodstadt, L., Nicod, J., Bhomra, A., Hernandez-Pliego, P., Whitley, H., Cleak, J., Dutton, R., Janowitz, D., Mott, R., Adams, D. J., and Flint, J. (2011). Sequence-based characterization of structural variation in the mouse genome. *Nature*, 477(7364) :326–9. (Page 123.)
- Zeitlinger, J., Zinzen, R. P., Stark, A., Kellis, M., Zhang, H., Young, R. A., and Levine, M. (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev*, 21(4) :385–90. (Page 47.)
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9) :R137. (Page 34.)

- Zhao, Y., Granas, D., and Stormo, G. D. (2009). Inferring binding energies from selected binding sites. *PLoS Comput Biol*, 5(12) :e1000590. (Page [24](#).)
- Zhou, Q. and Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6) :909–916. (Page [69](#).)
- Zhou, Q. and Wong, W. H. (2004). CisModule : de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A*, 101(33) :12114–9. (Page [52](#).)
- Zinzen, R., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269) :65–70. (Pages [45](#), [106](#) et [112](#).)
- Zykovich, A., Korf, I., and Segal, D. J. (2009). Bind-n-Seq : high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic Acids Res*, 37(22) :e151. (Page [31](#).)